Cognitive complexity, expertise and prediction accuracy in venture selection

Thomas Åstebro[a], Andrew Funck[a], Francesco Giordano[a]

[a] *HEC Paris, Economics and Decision Sciences*

Abstract

Venture capitalists, business angels, funding agencies, and incubators evaluate ventures, a difficult task where decision uncertainty is high. We examine how the degree of cognitive complexity and human expertise affects judges' admission recommendations at an incubator. Judges read an application, use preset criteria to score it, and form an intuitive overall judgment to accept or reject the application. We model and test how cognitive complexity and judge expertise affect this judgment through a Bayesian classification model, with implications on classification uncertainty and accuracy. Judges demonstrate poor accuracy in evaluating venture quality, but we show that the decision environment is so noisy that they can't do much better. A Bayesian model of judgment capture much of the decision variation. Complexity raises uncertainty and lowers classification accuracy, while expertise reduces uncertainty and improves accuracy only for moderately complex cases.

⋆We would like to express our sincere gratitude to those who provided comments on earlier versions: Jose Penalva, Carlos Serrano, Frederic Koessler, Brian Hill, Stefania Minardi, Itzhak Gilboa, Emmanuel Kemel, and Gilles Stoltz, as well as seminar participants at the 2nd Bayesian Entrepreneurship Conference in Milan, 3rd Workshop on AI and Entrepreneurship at HEC Paris, MSI Munich, Barcelona Summer School, Workshop on Crime and Entrepreneurship in Siena, Sumantra Ghoshal Conference, Nordic Centre for Entrepreneurship Research, IE Madrid, University Carlos III Madrid, Max Planck Institute for Innovation and Competition, Bologna University, and Bilkent University. This research was supported by the ION Management Science Lab – HEC Paris. Francesco Giordano has benefited from a State grant managed by the Agence Nationale de la Recherche under the Investissements d'Avenir programme with the reference ANR-18-EURE-0005 / EUR DATA EFM.

Email addresses: astebro@hec.fr (Thomas Åstebro), andrew.funck@hec.edu (Andrew Funck), francesco.giordano@hec.edu (Francesco Giordano)

1. Introduction

Selecting high–potential ventures is a first order decision in entrepreneurial ecosystems: early–stage resources are scarce, and their allocation by incubators, accelerators, funding agencies and investors shapes which ideas scale and which do not (Cohen and Hochberg, 2014; Kerr and Nanda, 2015).[1] Assessing venture quality, however, is inherently difficult: historical data are often sparse, reliable accounting indicators are typically unavailable, and evaluations must be made under substantial uncertainty about future outcomes. Hence, in one business plan competition in Nigeria, the average scores by a team of trained expert judges did not predict business success of the applicants three years later (McKenzie and Sansone, 2019). Nevertheless, research shows that some judges apparently can predict with high accuracy whether an invention will succeed reaching the market or not (Åstebro and Elhedhli, 2006), that judge evaluations in pitch competitions predict subsequent venture financing (Howell, 2020, 2021), that venture evaluators can meaningfully distinguish some high-quality ventures (deep tech) but not others (consumer products, SaaS) based on a short abstract (Scott et al., 2020), that "disagreement" – the dispersion in judges' evaluations in pitch competitions – predicts subsequent venture success (Gius, 2025), and that more diverse evaluation panels provide more informative predictions about venture outcomes than homogeneous panels, but only for above-average ventures (Lane and Rietzler, 2026). Apparently, judges may or may not play an important role in sorting good from bad ventures.

To explain what might be going on, we rely on the large literature in judgment and decision–making which shows that unaided human integration of multiple signals tends to be noisy and humans therefore often perform poorly in prediction tasks, frequently underperforming even simple actuarial models (Dawes et al., 1989; Grove and Meehl, 1996). Conditional on knowing the sign, linear additive models with any random weights do better than humans for two statistical reasons: first most models have flat likelihoods, meaning that any combination of weights does about equally well; second, human are unreliable and make errors given the

[1] Ventures refer to businesses in the process of being created.

same data but machines do not (Dawes, 1979). Dawes and Corrigan (1974), p. 105 conclude that "the whole trick is to know which variable to look at and then know how to add". (Kahneman, 2018), p. 610, is more sceptical and conclude that "one of the major limitations on human performance is not bias it is just noise" and, therefore, "you should replace humans by algorithms whenever possible... Humans do so poorly and are so noisy that just by removing the noise you can do better than people." As an example from venture judgment, (Zacharakis and Shepherd, 2001; Zacharakis and Meyer, 2000) show that a linear additive statistical model estimated on the classification decision of VCs venture performance predictions perform much better (60% accuracy) than the actual decisions made by the same VCs (39.5% accuracy). We are left with wondering whether unaided human judgment of ventures has any predictive properties, especially since most articles on venture judgments have investigated correlations rather than prediction accuracy.

What may cause prediction errors by venture judges? Experiments show that higher cognitive complexity of the decision environment increase decision errors in experimental situations (e.g. (Butler and Loomes, 2007; Enke and Graeber, 2023; de Clippel et al., 2024). Increases in cognitive complexity can come in many forms, such as having to compute compound rather than simple probabilities, higher cardinality of scales, and in more and a more diverse set of parameters with scales that are not directly comparable (Dawes, 1979; Enke, 2024). Cognitive complexity is likely to be high when assessing ventures as judges must typically infer underlying venture quality from many weak, noisy cues that are all differently scaled, and where an objective benchmark is not available. As an example, in the experiment by (Zacharakis and Shepherd, 2001; Zacharakis and Meyer, 2000), providing more predictive information to a VC decreased her prediction accuracy from 39.5% to 30.9%, a decrease which could be explained by an increase in the cognitive complexity. On the other hand, through both developing experience and training, humans can develop better prediction accuracy, although the benefits of training appear not to be that large (Moore et al., 2017; DellaVigna and Pope, 2018).

The question that we address is: how well does venture judges do in predicting venture success

as compared to an optimal use of information and under which conditions can these judges' prediction accuracy improve? The answer is of great interest for those selecting judges and designing venture evaluation processes, as well as for the accumulation of knowledge regarding how to best support ventures.

In this paper we use statistical decision theory to estimate a normative benchmark for how accurate venture judges may be: given (possibly noisy) evidence about a venture, the optimal evaluation is to compute the posterior probability of each quality state conditional on that evidence and assign the label with the highest posterior – that is, follow Bayes' rule. The maximum achievable accuracy is limited by the informativeness of the observable evidence; the Bayes classifier attains this information frontier (Berger, 1985; Devroye et al., 2013).

We apply this Bayesian framework to the selection process of an accelerator program funded by a leading business school in France, the HEC Incubateur. We model evaluators as Bayesian decision makers who combine a prior over venture quality with a likelihood implied by the observed evidence to form a posterior and then issue a binary recommendation. Empirically, we estimate an empirical Bayes classifier from the observed multidimensional scoring by the judges and use it as a normative benchmark for what can be achieved with the available evidence. Bayesian updating models are increasingly adopted to describe also entrepreneurial decisions even though they are typically made under ambiguity (Cohen and Koning, 2024; Nanda, 2024; Chavda et al., 2024). Although an established trail of compatriot papers use judges' scores in various ways (Åstebro and Elhedhli, 2006; Howell, 2020, 2021; Scott et al., 2020; Gius, 2025; Lane and Rietzler, 2026), our specific use of the judges' scores is new. Our data provide Granger causality. Similar to in some pitch competitions (Howell, 2020, 2021), the scoring and decisions to recommend a venture to be admitted or not by a specific judge is never revealed to the ventures, meaning that there is no mechanical causal effect of the score or judgment by a judge on venture success. However, ventures that are ultimately selected for inclusion in the program by the HEC Incubateur are likely to obtain a treatment effect, which we therefore control for.

We find that judges accurately identify the economic quality of ventures in approximately 61% of cases, and the Bayesian model performs similarly (62%). Random guessing (e.g., flipping a coin) would instead yield an accuracy of 50%. These levels indicate that both human and Bayesian decisions are only moderately accurate, which we conclude must reflect the limited informativeness of the available evidence. We then assess alignment with the model benchmark and find that 79% of individual recommendations coincide with the empirical Bayes classifications under symmetric misclassification costs; allowing for asymmetric costs that penalize rejecting good ventures more than admitting bad ones increases this alignment up to 85%. This pattern suggests that judges' decisions generally align with an optimal probabilistic classification rule, while limited accuracy derives primarily from the difficulty of extracting valuable information from the application dossier.

To explain how decision quality varies across judges and ventures, we propose a simple model that treats evidence as a noisy signal whose informativeness depends on two primitives, sources of cognitive uncertainty: evaluators' expertise and cognitive complexity. Cognitive complexity stem from the inherent difficulty of understanding an application. In the model, expertise increases the separation between the signal distributions for high– and low–quality ventures (making evidence more informative), whereas complexity compresses that separation (making evidence less informative). A judge who applies Bayes' rule is therefore more accurate when more expert and less accurate when the venture is more complex to understand; moreover, the return to expertise is largest at intermediate complexity, where there is something to learn but the task is not hopeless. These predictions translate into reduced–form hypothesis we can test in the data.

We measure cognitive complexity by either the application text complexity or the variation in the interpretation of the application across judges, as measured through the variation of a signal obtained by aggregating a set of noisy covariates (either free text or through a fixed set of constructs). Further, we measure judge expertise alternatively by a judge fixed effect representing some unobservable ability, or by the observable cumulative number of evaluations

a judge has performed prior to the next judgment. We use two measurable outcomes to test hypotheses; the accuracy in making correct classifications by a judge, and the remaining uncertainty that a judge retains after making a judgment. We test the robustness of results using a number of alternate measures and model specifications. We document that (i) posterior uncertainty is lower for more experienced judges and higher for more complex ventures, and; (ii) accuracy in identifying high-quality ventures decreases with complexity and does not improve as a given judge gains experience; yet, judges who perform more evaluations are more accurate on average, consistent with sorting on ability rather than learning-by-doing. Expertise is found to be most valuable for moderately complex cases.

2. Related Literature

Experiments show that people seem to make judgments that are consistent with models where "noise" affect the quality and type of decisions.[2] We use the term "cognitive complexity" to mean the noisiness of the decision environment. In our empirical setting we measures this, alternatively by the text complexity of the submitted application, and by the degree of disagreement across judges in a venture's score. Higher cognitive complexity appear to increase decision errors across many experimental situations (Butler and Loomes, 2007; Enke and Graeber, 2023; de Clippel et al., 2024). There are further implications of cognitive complexity for decision making that we are not able to cover in this study. For reviews of the experimental literature on cognitive complexity, see Enke (2024) and Oprea (2024).[3]

[2]For a generalist view on the role of noise in judgments and decision making, see Kahneman et al. (2021).

[3]To name a few examples, higher cognitive complexity is disliked (Oprea, 2020), and people therefore tend to reduce the complexity of the decision situation. de Clippel et al. (2024) show that people undervalue options they find complex; Halevy and Mayraz (2024) find that people prefer simple rules that they themselves design rather than optimal even though they are not less cognitively costly; Arrieta and Nielsen (2024) show that individuals' choice processes are more describable in complex choice environments, suggesting that they then tend to substitute complex evaluations with procedures, and when Banovetz and Oprea (2023) artificially remove complex calculations by having a computer track and organize past events, people abandon simpler rules and use maximally complex optimal rules instead. Higher cognitive complexity tend to attenuate decision parameters such as probabilities (Enke and Graeber, 2023) and valuations (de Clippel et al., 2024). And when forming predictive mental models of cognitively complex data, subjects (i) often fail, (ii) often cannot explicitly describe the model they have formed even when successful, and (iii) tend to be attracted to the simplest model (Kendall and Oprea, 2024). These articles represent only a small sample of the recent groundswell of experiments on cognitive complexity. In addition, a comprehensive review of the literature on humans as biased Bayesians can be found in Benjamin (2019). A review of recent economic models to take these biases

Not all researchers agree that the laboratory can well represent decision errors made in natural settings. For example, Schwartz and Griffin (2012) conclude that "decision heuristics... appear more likely to create biases in the psychology laboratory than in the [medical] clinic", and "biases found readily in other research are not evident in the judgements of professional auditors" (Smith and Kida, 1991). The laboratory is indeed created to form a precise but artificial situation. While field data may lack the precision of the laboratory it instead illustrates how decisions are made "as they appear." There is then an argument to take the models and empirical findings derived from many laboratory studies to test in the field. However, unfortunately only a small set of studies, limited to sports or stock trading data, show how humans perform in the field when facing cognitive complexity. Nevertheless, these studies support the lab experimental results reported above (Augenblick et al., 2025; Drerup et al., 2017; Giglio et al., 2021; Molavi et al., 2023).

People may face many challenges in making accurate judgments in natural complex decision contexts, including being required to integrate numerous pieces of information, a process riddled with human error (Dawes, 1975; Enke, 2024; Molavi et al., 2023)[4]. In the field, and in particular as it applies to assessing ventures, people, in addition, typically lack timely feedback on their judgments (Fischhoff, 1975; Åstebro and Koehler, 2007), or only obtain partial or biased feedback (Åstebro and Koehler, 2007). Feedback on the accuracy of past decisions is a precondition for learning (Dahlin et al., 2018). These field conditions are therefore likely to produce even more noisy venture judgments than in the lab.[5]

into account in a Bayesian updating framework is provided by Ortoleva (2022).

[4]There is a substantial literature in behavioral economics on the types of decision errors a person may have and how they might be detected, and potentially corrected. The decision-maker must first a) form an understanding of the opportunity set and the mapping from choices to distributions of consequences, b) formulate objectives, and c) optimize. Errors can arise in any of these three stages. For a discussion of some of the literature addressing error detection in these stages, see Bernheim et al. (2026). It is not our intent to decompose the various sources of errors that the judges face, but to provide a simple Bayesian updating model of the prediction problem, and then analyze how judges compare to this model. For supporting theory see for example Aragones et al. (2005).

[5]On the other hand, when feedback on decision in natural environments is quick and accurate, such as in weather forecasting, humans making probabilistic forecasts appear to be well calibrated along the entire probability distribution (Murphy and Winkler, 1984).

One potential factor that may diminish noise in judgments is learning-by-doing. Stated simply "As experience accumulates, someone-the manager, the worker, the engineer, the head of purchasing-makes better decisions." (Jovanovic and Nyarko 1995, p. 248). A number of articles show these empirically replicable patterns with steep efficiency improvements early on across different settings (e.g. Waldman et al. 2003; Mazur and Hastie 1978). Thompson (2010) and Dahlin et al. (2018) summarize the literature. Our study on the effects of judge learning on prediction accuracy and uncertainty is closely aligned with this literature.

Regarding the estimation of Bayesian updating models, Grether (1980), models and estimates that individuals often neglect priors when forming judgments, and to a lesser extent also discount information in the likelihood. Benjamin (2019) summarize the literature. We follow standard theory and do not assign any particular weights reflecting biased Bayesian updating. Our data are framed in a context where the HEC Incubator admits some ventures, with the premise of helping them to accelerate their path to success. This will, potentially, produce a treatment effect on outcomes for those admitted. For examples of such effects, see Gonzalez-Uribe and Leatherbee (2018); Cohen et al. (2019b); Yu (2020); Hallen et al. (2020); González-Uribe and Reyes (2021); Assenova and Amit (2024). We use established methods to control for such treatment effects in this setting, as outlined in McKenzie and Sansone (2019). We do not consider other aspects of the venture accelerator process that happens after the decisions we document in this article. For related work that examine the venture acceleration process in Incubators and Accelerators see for example Cohen et al. (2019a); Sharapov and Dahlander (2025), and Avnimelech et al. (2025). For related work on the judgment of science projects by committees see for example Kaplan et al. (2008), Criscuolo et al. (2017), Franzoni et al. (2025), and Lane et al. (2022). For related experimental work on differences between experts (or more cognitively high performing individuals) and lay people in prediction accuracy, see for example Moore et al. (2017); DellaVigna and Pope (2018); Chu et al. (2024); D'acunto et al. (2023).

Our analytical approach is grounded in the selection problem of the incubator. Our paper

is different than others examining the judgment of the inherent quality of ventures in that we focus on judges' prediction accuracy. That means that as opposed to establishing how much variance in the target is explained by a score, we model how the judgment to accept or reject a venture is determined by a judge's score, and how well that judgment performs. The process then involves determining how to construct the score, calibrating a score cut-off value for acceptance, and taking into account the relative cost of making two classification errors. Hence, for example, calibration can be high even when R2 is low (even non-existent) when class imbalance is high, simply by making an extreme decision on all applicants, but that would serve no practical purpose. On the other hand, when R2 of a regression is low, it is unlikely that calibration accuracy will be high even if coefficients are significant. (Gius, 2025) is close to us as he uses cross-judge variation in a score of judged venture quality to predict venture success. We use a similar construct of judge disagreement to instead examine its correlation with a) prediction accuracy, and b) the remaining uncertainty in the decision once it is made. (Gius, 2025) further use a text-based measure from the application package submitted to the venture competition to measure venture distinctiveness. We instead used a text based measures of the complexity of the application text to verify the results on judge disagreement, arguing that both measures are related but distinct measures of venture complexity. Åstebro and Elhedhli (2006) is the closest to us in modeling and estimating judge prediction accuracy, although they focus on extracting the decision heuristic used by the judges. Others are more concerned with correlating forecasts with venture success (Howell, 2021; Scott et al., 2020).

3. Context

The HEC Incubateur is a venture acceleration program based in Paris. It caters primarily to young entrepreneurs living in Paris who have recently graduated from university, often from an entrepreneurship program at HEC, or from its partner institutions and universities in Paris. Each quarter (referred to as a batch), the incubator admits approximately 10 ventures, depending on the availability of slots and the quality of applications received.

In the initial step of the admission process, applicants complete an online form. Some ap-

plicants are deemed ineligible and are automatically rejected. Those meeting the eligibility requirements proceed to the next stage where they are invited to complete a second application form. The application asks for information about the team, venture development stage, customers, market, business model, competition, product development, plans, funding, burn rate, regulatory concerns, and impact. Data regarding all eligible applications from 9 batches between Fall 2021 and Winter 2024 are shown in Table 1, covering 1,638 evaluations and 579 applicants.

– INSERT TABLE 1 ABOUT HERE –

40% of eligible applicant ventures have at least one HEC Paris graduate among their founders, while 46% include at least one female co-founder. The average founder age is 33 years, and teams typically consist of four members, with two co-founders. Additionally, 64% of ventures are incorporated at the time of application. Subscription-based businesses (54%) are most common, followed by marketplaces (21%) and e-commerce ventures (15%). Applicants span a broad set of sectors most frequently Software IT (23.7%), Finance/Real Estate (12.4%), and Life Sciences/Healthcare (11.4%) alongside Hospitality/Education and Consulting/Professional Services. Venture maturity also varies substantially: 47% report a "Minimum Viable Product" (a functional but early-stage version), 27% are still in the prototype phase, and 26% have a fully developed product at the time of application.

Each application is assessed by at least two professional judges who are randomly assigned from a pool of volunteers. Judges include former entrepreneurs, investors, academics, and incubator staff. For each batch, judges indicate their availability, and the incubator then randomly assigns them to a set of startups. Judges are not informed of the identity of the other judges evaluating the same application and assess ventures independently on fourteen criteria, providing both criterion-level grades and a trinary admission recommendation. The criteria cover, among other dimensions, how clearly a user pain point is articulated, whether the proposed solution is feasible, the venture's competitive advantage, the founders' ability to execute, and the project's financial viability. Interviews with judges suggest that admitting

high-quality ventures is their primary concern. Secondary considerations, such as maximizing incubator impact and maintaining a balanced portfolio, are generally not incorporated at this stage, since judges lack full information about the broader applicant pool.

For each application, a judge assigns a grade ranging from 1 (lowest) to 3 (highest) for each criteria. Criteria grades measure application quality and serve as key indicators for judges when making recommendations on a venture. At the end of the form, a judge is required to indicate whether she recommends the application to proceed to the next stage of the admission process, again using a scale from 1 (No) to 3 (Yes), with 2 being uncertain. Across all batches, judges provide positive recommendations in about 40% of evaluations.

There are two additional stages leading to an admission or rejection. They contain a committee meeting among a subset of the staff judges that made the initial assessments, and a one-day interview stage by an external Jury for a final group of applicants. In these two meetings, participating judges get to hear and see what some of the other judges thought. Feedback on earlier judgments are therefore available but only for those participating. These stages, however, contain different judgment processes and considerations and are therefore not part of the analysis in this paper. For example, strategic considerations are discussed in the committee meeting. While between 35% and 46% ventures in each batch are recommended to be accepted in the first stage, approximately 26% to 46% pass the second stage and 11% to 25% ventures are ultimately accepted by the final Jury.

We collected data on the economic performance of all applicants during the summer of 2024. Using the full name of the main applicant (typically the CEO), the co-applicant, and the business name provided in the application, we manually searched for information on each venture. Data were gathered from multiple sources, including LinkedIn, Crunchbase, and company or personal websites. The performance indicators include the number of full-time employees (FTEs), total funding raised, and firm survival status. Following established practice (Lerner and Malmendier, 2013; Howell, 2020), we assume that a business has ceased operations when no information is available from any of the three data sources and assign it a value of zero for

survival.

– INSERT TABLE 2 ABOUT HERE –

Table 2 presents summary statistics on ventures' applications, survival rates, FTEs, and funding outcomes, both by batch and in aggregate. Each batch includes approximately 50 to 100 evaluated applications. As of summer 2024, the share of ventures still active ranges from 56% to 85%, with an average survival rate of 73%. Among surviving ventures, the average number of FTEs is 6.6, with the 75th percentile reaching 8. Regarding funding outcomes, an average of 11% of ventures obtain post-application funding. Among those that do, the mean amount raised is €3.5 million. This figure is skewed upward by a small number of ventures that secured exceptionally large investments, particularly in the earlier batches.

This setting describes a decision environment in which judges must form binary admission recommendations under substantial uncertainty, based on noisy and multidimensional information, and with limited opportunities for feedback or learning from outcomes. Judges differ in their experience with the task, while ventures differ in how difficult they are to assess. These features motivate the conceptual framework developed in the next section, which models venture evaluation as a probabilistic classification problem in which judges combine prior beliefs with imperfect signals of venture quality. The framework formalizes how evaluator expertise and cognitive complexity jointly shape the precision of these signals, and thus determine both decision accuracy and residual cognitive uncertainty.

4. Conceptual framework

We propose a model of the recommendation process of judges. Judges possess a level of expertise that may depend on their in-task experience (e.g., the number of prior evaluations they have conducted) as well as their task-related external experience (e.g., their professional background or domain of expertise). Each judge evaluates a venture characterized by a specific level of evaluation complexity. By performing an evaluation, a judge produces a subjective signal whose precision depends jointly on the judge's expertise and on the difficulty of eval-

uating the venture. Based on the observed signal, the judge then issues an acceptance or rejection recommendation according to a decision rule that assigns, to each possible signal realization, a probability of recommending acceptance. All technical details, derivations and proofs are provided in Appendix I, together with a more general model that relaxes many of the assumptions postulated in this section.

The evaluation model, which represents the mechanism underlying signal generation, characterizes the distribution of signals produced by a judge with a given level of expertise when assessing a venture of a given complexity, conditional on the venture's unobserved true value. This formulation reflects the judge's evaluation skills, capacity to acquire and process information (e.g., from application data), and private knowledge. Two key performance measures are of primary interest. The first is decision accuracy, defined as the rate of correct decisions achieved under an optimal decision rule given the evaluation model. The interpretation of optimal accuracy concerns the maximum attainable level of correct recommendations a judge can make, conditional on their inference capacity. The second measure is cognitive uncertainty, the degree of uncertainty a judge retains after processing information about a venture's value. Higher cognitive uncertainty corresponds to greater variability in the judge's perceived valuation of the venture.

For simplicity, we assume that a venture can be either good or bad, depending on whether its underlying economic quality meets the incubator's standards and capacity constraints. Defining the state space accordingly, we thus model the judge's selection problem as a binary prediction task, where the judge's decision must align with the venture's true underlying state. More specifically, each judge forms a subjective assessment by first acquiring information about the venture's economic quality – modeled as a noisy signal generated by a statistical experiment – and then inferring the venture's value from this signal. Judgments are thus modeled as outcomes of subjective evaluation processes, where the distribution of signals depends jointly on the judge's expertise and the venture's complexity. Figure 1 illustrates the judge's inference and decision-making process.

— INSERT FIGURE 1 ABOUT HERE —

More formally, a latent variable $w \sim \mathcal{N}(0,1)$ represents the economic quality of the venture. The venture is classified as good if w lies in the high p percentile, that is, if $w \geq \Phi^{-1}(1-p) = \bar{w}$. Hence, we denote by $v \in \{0,1\}$ a random variable representing the "good" (1) or "bad" (0) value of a startup and $p = \Pr(v=1)$ be the prior probability of a startup being good. A judge evaluating a certain venture produces a noisy signal $s \in \mathcal{S} \subseteq \mathbb{R}$ about the venture's economic quality

$$s = w + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \frac{\sigma}{e}\right), \tag{1}$$

where ϵ is an error term independent on the underlining economic quality. The symbols e and σ denote independent random variables that jointly determine the precision of the signal. Variation in precision may arise either from the venture's intrinsic characteristics, captured by a venture-specific parameter σ, or from the evaluator's expertise, represented by e. The posterior probability that a startup is of high quality, given a signal, is denoted by $\eta(s) = \Pr(v=1|s)$.

The noise of the signal – modeled as the product of the venture-specific and evaluator-specific components – captures some degree of complementarity between expertise and complexity. This framework represents the entire evaluation problem as a joint probability distribution, $p(s, v, w, e, \sigma)$, defined over the model primitives: the signal (s), the venture's binary label (v), its underlying quality (w), the judge's expertise (e), and the venture's idiosyncratic complexity (σ). The evaluation model, that is, the probability of observing a signal given a venture with value and idiosyncratic complexity level, evaluated by a judge with fixed expertise, is defined as

$$f(s|v,\sigma,e) = \frac{\int_w p(w)p(s|w,e,\sigma)p(v|w)dw}{\int_{s,w} p(w)p(s|w,e,\sigma)p(v|w)dsdw}.$$

4.1. Decision problem and objective function

An evaluator, given his signal, decides whether or not to recommend a venture. The decision is represented by a map $a : \mathcal{S} \to [0,1]$ that to each signal realization assigns a probability of admission recommendation. We consider the problem of selecting ventures that fall into the

high percentile of economic quality

$$\max_{a:\mathcal{S}\to[0,1]} \mathbb{E}_s\Big[a(s)v + \gamma(1-a(s))(1-v)\Big], \quad \text{with} \quad v = \mathbf{1}_{\{w\geq\bar{w}\}} \tag{2}$$

where γ reflect the ratio of the costs of missclassification. For simplicity and tractability of the analytical results, in this section we consider the case where $\bar{w} = 0$ and $\gamma = 1$, hence the incubator aims to select ventures that perform better than average and gives equal cost to selecting bad ventures and rejecting good ones. Both assumptions are without loss of generality, as argued in Appendix I. The optimal decision rule[6] depends on the expected venture quality conditional on the signal, which is directly related to the posterior probability of high quality, $\eta(s) = \Pr(v = 1 \mid s)$, and is given by

$$a^*(s) = \begin{cases} 1 & \text{if } \mathbb{E}[w|s] \geq \bar{w} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathbb{E}[w|s] = \frac{e}{e+\sigma}s$. Namely, judges should recommend a venture when the excepted quality given the signal is high enough.

The optimal expected accuracy depends on the precision of the signal – hence on the parameters σ and e. In this setting, the optimal accuracy[7] has an explicit analytical expression:

$$\begin{aligned} \xi(e,\sigma) &= \mathbb{E}_s\left[a^*(s)v + (1-a^*(s))(1-v)\right] \\ &= \frac{1}{2} + \frac{1}{\pi}\arctan\sqrt{\frac{e}{\sigma}}. \end{aligned} \tag{4}$$

Proposition 1 (Expected Accuracy)**.** The following statements hold:

1. The optimal expected accuracy is decreasing in cognitive complexity.

[6] For a formal derivation, see Appendix I. Intuitively, note that $\mathbb{E}[v \mid s] \geq \frac{1}{2}$ if and only if $\mathbb{E}[w \mid s] \geq \bar{w}$.
[7] For the analytical derivation the reader is referred to Appendix I.

2. The optimal expected accuracy is increasing in judge expertise.

Furthermore, the optimal expected accuracy is concave in the expertise level, suggesting, for example, that the cumulative number of evaluations has decreasing marginal returns in terms of classification accuracy, and is convex in cognitive complexity, indicating that the marginal impact of complexity on classification accuracy attenuates as complexity increases.

The marginal value of expertise across different levels of cognitive complexity can be examined by the difference $\xi(\sigma, \bar{e}) - \xi(\sigma, \underline{e})$ in the optimal expected accuracy for two values of expertise levels $\bar{e} \geq \underline{e}$, as a function of σ. This measure provides quantitative insight into the following questions: to what extent do more expert judges outperform less expert judges as the complexity of the underlying problem varies? For which levels of complexity is expertise most valuable? More generally, the marginal value of expertise can be studied through the partial derivative, $\xi_e(\sigma, e)$, of the optimal expected accuracy with respect to the expertise level. In this specific setting, the marginal value of expertise is tractable analytically and given by

$$\xi_e(\sigma, e) = \frac{1}{2\pi} \frac{\sqrt{\sigma}}{\sqrt{e}} \frac{1}{e + \sigma}. \tag{5}$$

Equation 5 suggests that the difference between more and less expert judges is maximized at some intermediate values of complexity[8], hence, expertise is more valuable at intermediate complexity levels. In parallel, for very simple and very complex problems, novices tend to do as well as more experienced judges.

While accuracy captures expected classification performance, it does not describe how much uncertainty remains about a venture's value after observing the signal. We therefore complement accuracy with a measure of cognitive uncertainty, defined as the dispersion of posterior beliefs. A natural measure of cognitive uncertainty is the expected variance of a venture's value under the posterior distribution. Qualitatively, this measure reflects how uncertain is

[8] see Figure K3 in the Appendix K

a judge about the true economic value of a venture after receiving some information. The cognitive uncertainty of a judge with expertise level proportional to e, evaluating a company with idiosyncratic noise σ, is thus

$$\mathbb{E}_s[\mathrm{Var}(v|s)].$$

cognitive complexity and judge expertise impact the cognitive uncertainty as predicted by the following proposition, which is derived in Appendix.I.

Proposition 2 (Cognitive Uncertainty)**.** The following statements hold:

1. Cognitive uncertainty increases with cognitive complexity.
2. Cognitive uncertainty decreases with judge expertise.

The conceptual framework leads to the formulation of a set of testable verbal hypotheses centered on two key dimensions: evaluator expertise and cognitive complexity. Expert judges are expected to provide more accurate evaluations than non-experts, with accuracy declining as cognitive complexity increases. Simpler ventures and more expert evaluators yield lower posterior variance, and the accuracy gap between experts and non-experts is greatest at intermediate levels of complexity.

Hypothesis 1a	*Expertise improves evaluation accuracy.*
Hypothesis 1b	*Expertise attenuates cognitive noise.*
Hypothesis 2a	*cognitive complexity deteriorates evaluation accuracy.*
Hypothesis 2b	*cognitive complexity increases cognitive noise.*
Hypothesis 3	*The accuracy gap between experts and non-experts grows with complexity up to an intermediate level, then declines.*

5. Empirical Strategy

This section develops the empirical framework used to evaluate empirically the conceptual model. We proceed in three steps. First, we define the empirical counterparts of each the-

oretical construct: venture quality, the evaluation signal, the judge's recommendation, judge expertise, and cognitive complexity (§5.1). Second, we describe the estimation of an empirical Bayes classifier to identify high-quality ventures based on the score from venture evaluations, replicating the decision process described in Section 4. (§5.2). Third, we present the regression specifications used to test the model's hypotheses regarding the roles of expertise and complexity in shaping decision accuracy and cognitive uncertainty (§5.3).

5.1. Measurements and empirical counterparts of model components

Each theoretical construct is mapped to an observable empirical counterpart. We index ventures by i and judges by j. Where necessary, variables are defined at the evaluation level (i, j). We use the hat notation (e.g., $\hat{v}$) to distinguish empirical measures from their theoretical analogues.

Venture quality. The binary quality label $\hat{v}_i$ classifies each venture as high-quality ($\hat{v}_i = 1$) or low-quality ($\hat{v}_i = 0$) on the basis of ex-post economic performance. We measure performance using two indicators observed as of 2024: (i) the number of full-time employees (FTEs) and (ii) the cumulative amount of capital raised, including grants. To remove mechanical variation arising from differential exposure time across admission batches, we residualize both outcomes with respect to batch fixed effects. We then rank ventures separately on each residualized measure, construct a composite performance index as the equally weighted average of the two ranks, and set the index to zero for ventures that are no longer operational. The binary variable $\hat{v}_i$ equals one if venture i falls in the top 40% of this distribution and zero otherwise. The 40% threshold reflects the admission capacity of the incubator's first selection stage.

Ex-post performance may reflect not only underlying quality but also the causal effect of incubation for admitted ventures. We capture this potential treatment effect with a parameter δ, which applies only to admitted ventures and equals zero otherwise. Our baseline construction sets $\delta = 0$ for all ventures when mapping outcomes to underlying quality. In the robustness analysis (Section 7), we relax this assumption by recomputing classifications under alternative values of δ, allowing for plausible treatment effects on survival, employment, and capital

raised. Table L4 in the Appendix reports reclassification rates, which are modest across specifications, indicating that the quality labels are not driven by post-admission treatment effects.

Evaluation signal. The signal $\hat{s}_{i,j}$ is the score assigned by judge j to venture i, computed as the sum of fourteen criterion-level grades, representing an equal-weights linearly additive model. We do not know how each judge makes use of the fourteen pieces of information, but the sum (or the average) provides a simple aggregation of a multidimensional vector into a one-dimensional signal, representing a first-pass assumption about how a judge would him- or herself aggregate the information. (Recall the introductory discussion about human fallibility with respect to using more complex rules.) Other articles using similar information assumes the same aggregation model to form a score (Howell, 2020, 2021; Scott et al., 2020; Gius, 2025; Lane and Rietzler, 2026). Section 6.2 estimates the Bayesian model also using the standard signal structure in Bayesian classification (the "Naive Bayes"), which is the joint distribution of criterion level grades under the assumption of conditional independence. Figure K1 in the Appendix compares score distributions for high- and low-quality ventures. As expected, the distribution for high-quality ventures is shifted right to the distribution for low-quality ventures and exhibits lower variance, but both distributions overlap, consistent with interpreting $\hat{s}$ as an informative but imperfect signal of underlying economic quality.

Recommendation. The recommendation $\hat{a}_{i,j}(\hat{s}_{i,j})$ is a binary variable equal to one if judge j recommends venture i for admission after assigning score $\hat{s}_{i,j}$, and zero otherwise.

Judge expertise. We proxy judge expertise using evaluation activity. Specifically, $\hat{e}_{i,j}$ denotes the number of evaluations completed by judge j prior to evaluating venture i. The distribution of cumulative evaluations is highly skewed: the bottom 50% of judges account for only 10% of all evaluations, while the high 25% conduct more than 75% (Figure K2 in the Appendix). Alternatively we use a judge fixed effect, representing an unobservable level of expertise. Using a judge fixed effect to proxy an observable judgment attitude follows estab-

lished literature in various domains of law, insurance, and patent examination (Kling, 2006; Maestas et al., 2013; Dobbie and Song, 2015; Sampat and Williams, 2019).

Cognitive complexity. We capture cognitive complexity using two complementary measures. The first is based on judge disagreement: we compute the normalized within-venture variance of the residuals from a regression of scores $\hat{s}_{i,j}$ on judge fixed effects. This procedure isolates idiosyncratic disagreement across judges evaluating the same venture from systematic differences in grading tendencies, and thus measures how mentally difficult a venture is to assess. The theory of cognition derived from this measure is as follows. Single item measurements may contain a lot of noise. But as information is aggregated across a relatively large number of measures (14) into the score, large differences in this score across judges must mean something. If the same application is judged very differently by different judges as evidenced by this score, we infer that the application must therefore be more cognitively complex to judge. The second measure is text-based and more directly measures cognitive complexity when reading a text: the RIX readability index computed from the application text of each venture i (see (Anderson, 1983)). RIX captures linguistic complexity—sentence length and vocabulary demand— and therefore reflects how difficult the application is to process mentally based on its textual content alone, independently of judge realized disagreement. We therefore expect these two measures to capture different dimensions of cognitive complexity.

5.2. Empirical Bayes Classifier

The empirical Bayes classifier is estimated and assessed using a leave-one-out procedure. For each venture in the sample, the model is trained on evaluations of all remaining ventures and then assessed on evaluations of the held-out venture. Every venture serves exactly once as the held-out observation.

The unit of analysis is the individual evaluation. The estimation proceeds in two steps. In the first step, we estimate the score distribution conditional on ex-post venture type—the density of scores among evaluations of high-quality ventures, $\hat{f}_{\hat{S}|\hat{v}=1}$, and among evaluations of low-quality ventures, $\hat{f}_{\hat{S}|\hat{v}=0}$—using kernel density estimation with a linear kernel and bandwidth

of 0.32. To increase precision, we pool evaluations across judges and batches, thereby imposing a common conditional score distribution over judges and time.[9]

In the second step, for each evaluation of the held-out venture, we combine these estimated likelihoods with the prior $\hat{p} = 0.40$ to compute the posterior probability that a given score corresponds to a high-quality venture via Bayes' rule:

$$\begin{aligned} \hat{\eta}(\hat{s}_{i,j}) &= \hat{\Pr}(\hat{v} = 1 \mid \hat{S} = \hat{s}_{i,j}) \\ &= \frac{\hat{f}_{\hat{S}|\hat{v}=1}(\hat{s}_{i,j})\,\hat{p}}{\hat{f}_{\hat{S}|\hat{v}=1}(\hat{s}_{i,j})\,\hat{p} + \hat{f}_{\hat{S}|\hat{v}=0}(\hat{s}_{i,j})\,(1-\hat{p})}. \end{aligned} \tag{6}$$

Following the optimal decision rule derived in Section 4, we classify an evaluation as high-quality whenever $\hat{\eta}(\hat{s}_{i,j}) \geq 1/2$, which corresponds to equal weighting of Type I and Type II errors under zero–one loss. We denote the resulting classifier by $C_{\hat{\eta}} := \hat{a}^*(\hat{s}_{i,j})$.

5.3. Model predictions

We test Hypotheses **HP 1a**,**HP2a** and **HP3** using regression analyses that relate the accuracy of each recommendation, $\text{Accuracy}_{i,j} = \mathbf{1}_{\{\hat{a}_{i,j}=\hat{v}_i\}}$, to measures of judge expertise and cognitive complexity.

Expertise and accuracy (HP1a). To examine whether more experienced judges achieve higher accuracy, we estimate a correlated random effects model:

$$\begin{aligned} \text{Accuracy}_{i,j} = \alpha &+ \beta_1 \left(\log(\hat{e}_{i,j}) - \overline{\log(\hat{e}_j)}\right) \\ &+ \beta_2 \overline{\log(\hat{e}_j)} + \beta_3 \hat{s}_{i,j} \\ &+ \delta_{t(i)} + \theta_i + \gamma_j + \varepsilon_{i,j}, \end{aligned} \tag{7}$$

where $\log \hat{e}_{i,j}$ denotes the log number of evaluations previously completed by judge j before evaluating venture i, and $\overline{\log \hat{e}_j}$ is judge j's average log experience across evaluations. The

[9] This pooling assumption is convenient but potentially restrictive given heterogeneity in judging behavior and in the composition of ventures across batches.

log form represents typical learning functions (Thompson, 2010). The coefficient β_1 captures the within-judge relationship between experience and accuracy: it measures whether a given judge becomes more accurate as she accumulates evaluations relative to her own average experience level. The coefficient β_2 captures the between-judge relationship: it measures whether judges who are more experienced on average are also more accurate on average. Controlling for the score $\hat{s}_{i,j}$ allows us to compare judges facing the same observed evaluation signal. Batch fixed effects ($\delta_{t(i)}$) absorb common differences across admission rounds. Venture fixed effects (θ_i) further restrict identification to comparisons across judges evaluating the same venture, removing heterogeneity in evaluation difficulty. Judge fixed effects (γ_j) absorb all time-invariant judge heterogeneity and represents a within-judge unobservable level of expertise. Standard errors are clustered at the venture and judge level, not just in this but in all regressions.

Cognitive complexity and accuracy (HP2a). To test whether cognitive complexity reduces accuracy, we estimate:

$$\text{Accuracy}_{i,j} = \alpha + \sum_{f=2}^{4} \beta_f \hat{\sigma}_i^f + \hat{s}_{i,j} + \delta_{t(i)} + \gamma_j + \varepsilon_{i,j} \tag{8}$$

where $\hat{\sigma}_i^f$ are indicators for quartiles of cognitive complexity, $\hat{s}_{i,j}$ is the evaluation score, $\delta_{t(i)}$ are batch fixed effects, and γ_j are judge fixed effects. The coefficients β_f capture how recommendation accuracy varies with cognitive complexity relative to the omitted lowest-complexity quartile. We estimate this specification using both the disagreement-based and the text-based measures of complexity. Judge fixed effects ensure that observed accuracy differences are attributed to cognitive complexity rather than to systematic differences across judges.

Interaction effects (HP3). To assess whether the accuracy premium for experience varies

with cognitive complexity, we estimate:

$$\begin{aligned}\text{Accuracy}_{i,j} = \alpha &+ \beta_1 \text{High Experience}_j \\ &+ \sum_{f=2}^{4} \beta_f \hat{\sigma}_i^f \\ &+ \sum_{f=2}^{4} \gamma_f \text{High Experience}_j \hat{\sigma}_i^f \\ &+ \hat{s}_{i,j} + \delta_{t(i)} + \varepsilon_{i,j}\end{aligned} \tag{9}$$

where $\text{HighExperience}_j = 1$ for judges above the median of the cumulative evaluation distribution. The interaction coefficients γ_f test whether the accuracy gap between experienced and inexperienced judges is largest at intermediate complexity levels, as predicted by the model.

Cognitive uncertainty (HP1b, HP2b). To examine how cognitive uncertainty varies with expertise and complexity, we first allow the Bayesian updating process to differ across evaluation environments by estimating separate empirical Bayes classifiers on subsamples defined by median splits of judge experience and cognitive complexity. Estimating separate classifiers by subgroup allows the conditional score densities, and therefore the posterior belief function, to vary with both expertise and complexity. This, in turn, allows us to assess how the informativeness of evaluation signals differs across judges and ventures. Using the classifier estimated for the relevant subsample, we compute for each evaluation (i, j) the posterior variance

$$V(\hat{v}_i \mid \hat{s}_{i,j}) \;=\; \hat{\eta}(\hat{s}_{i,j})\big(1 - \hat{\eta}(\hat{s}_{i,j})\big), \tag{10}$$

which measures the cognitive uncertainty about venture quality after observing the score assigned by judge j. Higher values indicate that the score is less informative about whether the venture belongs to the high-quality class.

We then relate cognitive uncertainty to judge experience and cognitive complexity in separate

regressions. To study experience, we estimate:

$$\begin{aligned} V(\hat{v}_i \mid \hat{s}_{i,j}) = \alpha &+ \beta_1 \text{High Experience}_j \\ &+ \beta_2 \hat{s}_{i,j} + \beta_3 \hat{s}_{i,j}^2 \\ &+ \delta_{t(i)} + \theta_i + \varepsilon_{i,j}, \end{aligned} \tag{11}$$

where HighExperience_j is an indicator for judges above the median of cumulative evaluations. Venture fixed effects absorb all time-invariant heterogeneity in the evaluated venture, so β_1 captures whether, conditional on the score assigned, more experienced judges generate signals associated with lower cognitive uncertainty.

To study complexity, we estimate:

$$\begin{aligned} V(\hat{v}_i \mid \hat{s}_{i,j}) = \alpha &+ \beta_1 \text{High Complexity}_i \\ &+ \beta_2 \hat{s}_{i,j} + \beta_3 \hat{s}_{i,j}^2 \\ &+ \delta_{t(i)} + \gamma_j + \varepsilon_{i,j}, \end{aligned} \tag{12}$$

where HighComplexity_i is an indicator for ventures above the median of the complexity distribution, measured using either the disagreement-based or the text-based proxy. Judge fixed effects absorb time-invariant heterogeneity across judges, so β_1 captures whether more complex ventures generate signals associated with greater congnitive uncertainty, conditional on the score assigned. In both specifications, the score polynomial flexibly accounts for the mechanical relationship between cognitive uncertainty and the location of the score in the signal distribution, and batch fixed effects absorb common differences across admission rounds.

6. Results

6.1. Judges Classification Accuracy

How well do individual judges classify ventures with the highest economic value? Recommended ventures show significantly higher survival rates and employ more full-time equivalents than those rejected — on average, a 17 percentage-point higher survival rate, where the mean

is 73% and 3.9 more FTEs, where the mean number of FTEs is 6.7. Recommendations are less but still significantly associated with future funding outcomes, with a 0.5M difference at a mean of 3.5M (Table L1 in Appendix L). Overall, judges correctly classify 63% of evaluations, a figure that can be decomposed through the confusion matrix in Figure K4 in Appendix K: judges correctly recommend roughly 54% of high-quality ventures while correctly rejecting about 70% of low-quality ones.

To understand how this accuracy arises, we examine how judges translate scores into recommendations. Figure 2 plots recommendation rates against the actual probability of being high quality across score bins. Judges appear to follow an implicit threshold rule, recommending ventures once the posterior probability of high quality exceeds approximately 40%. Below this level, recommendations are rare; above it, they become the norm. This behavior is consistent with a Bayesian decision rule in which judges compare posterior beliefs to a cutoff.

– INSERT FIGURE 2 ABOUT HERE —

This threshold behavior has direct implications for the distribution of decision errors. Figure 3 decomposes evaluations within each score bin into true positives, false negatives, true negatives, and false positives. False negatives — high-quality ventures that are not recommended — concentrate at low and intermediate scores, while false positives cluster at higher scores. Overall accuracy, measured as the sum of true positive and true negative rates, is lowest in the intermediate score range where recommendation behavior switches most sharply. This pattern indicates that judges face the greatest cognitive uncertainty near the implicit decision threshold, and that the discrete nature of the recommendation amplifies the uncertainty inherent in the underlying signal.

– INSERT FIGURE 3 ABOUT HERE —

6.2. The Empirical Bayes classifier

We next assess how closely judges' behavior conforms to the Bayesian decision rule and how well this rule predicts venture quality. The estimated Bayes classifier $C_{\hat{\eta}}$ implements a deter-

ministic threshold on the score: ventures are classified as high-quality whenever the posterior probability $\hat{\eta}(\hat{s}) \geq \frac{1}{2}$, corresponding to equal misclassification costs.

Judges' recommendations align closely with this rule. Under symmetric costs, the classifier replicates 79% of judges' decisions. Allowing for asymmetric costs – consistent with judges' observed tendency to begin recommending at a posterior of roughly 40% rather than 50% – increases alignment to approximately 85%. Figure 4 visualizes this distinction by highlighting both the symmetric-cost admission region ($\hat{\eta}(\hat{s}) \geq 0.5$) and the asymmetric-cost admission region ($\hat{\eta}(\hat{s}) \geq 0.4$), showing that judges' observed decisions align more closely with the latter.

— INSERT FIGURE 4 ABOUT HERE —

Discrepancies between the classifier and judges' actual decisions are concentrated in the intermediate score range, where the posterior is close to the decision threshold and judges sometimes admit and sometimes reject ventures with similar scores, producing a U-shaped pattern in balanced accuracy across scores (Figure K5 in the Appendix).

Turning to predictive performance, we assess how well the classifier identifies high-quality ventures from scores. The empirical Bayes classifier attains a balanced accuracy of 60%, close to the benchmark accuracy of judges' actual recommendations (62%). A similar balanced accuracy (60%) is obtained using the standard Naive Bayes benchmark, which treats the full vector of criterion-level grades as the signal and assumes conditional independence across criteria. (We do not present further details of that analysis.) The Brier score (Murphy, 1973) is 0.233, a modest improvement over a constant base-rate forecast (0.244). A decomposition of the Brier score (Table L2 in the Appendix) shows that uncertainty is large, while resolution is limited: predicted posteriors remain clustered around the base rate, implying that the score only weakly separates high- from low-quality ventures (Figure K6 in the Appendix). By contrast, reliability is high: the expected calibration error is low (≈2.1%) and the calibration curve lies close to the 45-degree line (Figure K7 in the Appendix). Overall, the main constraint for the Bayes classifier is resolution rather than miscalibration.

The proximity of $\hat{\eta}(\hat{s})$ to $\eta(\hat{s})$ suggested by the low calibration error implies that the moderate accuracy of $C_{\hat{\eta}}$ reflects the limited informativeness of the score rather than estimation error in the posterior mapping. Through the distance between $\hat{\eta}(\hat{s})$ and $\eta(\hat{s})$, standard inequalities indeed bound the gap between the estimated classifier $C_{\hat{\eta}}$ and the theoretical classifier C_{η}, which thresholds the true posterior $\eta(\hat{s})$ and maximizes accuracy among all classifiers that are functions of the score $\hat{s}$ (Devroye et al., 2013).

6.3. Accuracy, cognitive uncertainty, expertise and complexity

Expertise and accuracy (HP1a). We next examine the relationship between judges' evaluation accuracy and experience. Table 3 reports estimates from specification (7), which decomposes experience into within-judge and between-judge components. The within-judge coefficient on log experience is negative and statistically significant across specifications. As a given judge accumulates additional evaluations over time, her accuracy declines. This result is robust to the inclusion of batch fixed effects (column 3), venture fixed effects (column 4), and judge fixed effects (column 5), the latter providing the most direct estimate of the within-judge learning component. The between-judge component points in the opposite direction. Judges who evaluate more ventures on average tend to be more accurate on average, with a positive coefficient that is marginally significant in the baseline specifications (columns 1–2). Once batch and venture fixed effects are included, the between-judge component is identified from substantially more limited residual variation across judges, which reduces statistical power and likely contributes to the loss of significance. Overall, these results clarify the relationship between expertise and accuracy in Hypothesis H_{1a}. Accuracy is positively associated with experience across judges, but within a given judge, additional experience over time is associated with lower performance. This pattern contradicts a learning-by-doing interpretation and is instead consistent with a selection mechanism whereby judges with higher underlying evaluative ability choose to perform more evaluations.

— INSERT TABLE 3 ABOUT HERE —

Cognitive complexity and accuracy (HP2a). Table 4 reports estimates using complex-

ity proxied by judges' disagreement on the same venture (columns 1–4) and by text-based measures of complexity of the application text (columns 5–8). Accuracy declines sharply with judges' disagreement in the score: relative to the least complex quartile, accuracy is lower by about 8–11 percentage points in the third quartile for both measures and by about 8–11 percentage points in the fourth quartile for disagreement. Estimates for the third quartile are statistically significant at the 1% or 5% level throughout, while the fourth quartile is statistically significant at the 1% or 5% level for disagreement. The second quartile shows a smaller and generally insignificant decline. The pattern is robust to including batch fixed effects, the score, and judge fixed effects. We examined also the correlation between the two measures of complexity and found it to be close to zero. That is, while both measures are capturing entirely different domains of application complexity, they show the same qualitative results. Overall, we conclude that higher cognitive complexity results in lower classification accuracy.

— INSERT TABLE 4 ABOUT HERE —

Interaction Effect (HP3). To assess whether experience matters more for more cognitively complex cases, figure 5 plots the accuracy gap between experienced and less experienced judges across quartiles of cognitive complexity, estimated from specification (5) using the text-based complexity measure. The gap is close to zero in the lowest complexity quartile, increases at intermediate levels of complexity, and then declines in the upper quartiles. The premium is largest in the second quartile, where experienced judges are about 12 percentage points more accurate. This pattern is consistent with the idea that expertise is most valuable for moderately complex cases: simple ventures can be classified accurately even by less experienced judges, while in very complex cases uncertainty limits the returns to any level of expertise.

— INSERT FIGURE 5 ABOUT HERE —

Appendix Table L3 reports the full set of interaction estimates underlying Figure 5 for both complexity measures. Using the text-based measure, the results are robust: the interaction between high experience and second-quartile complexity is positive, large (11.8–12.9 percent-

age points), and statistically significant at the 5% level across all three specifications, while the interactions at higher complexity quartiles are positive but not statistically significant. Using the disagreement-based measure, by contrast, the interaction coefficients are uniformly negative and generally not statistically significant, suggesting that the hump-shaped pattern documented in Figure 5 is specific to the text-based operationalization of complexity. Overall, the evidence provides partial support for the model prediction: the predicted non-monotonic interaction between expertise and complexity appears clearly when complexity is measured through textual features of the application but not when it is proxied by score disagreement.This suggests that the two measures capture different dimensions of complexity, which may interact differently with judges' experience.

Cognitive uncertainty (HP1b, HP2b).

Cognitive uncertainty results are aligned with theoretical expectations. Table 5 shows that the average cognitive uncertainty is lower for high-experience judges than for low-experience judges: specifically, judges with more than the median cumulative number of evaluations exhibit an average cognitive uncertainty that is 7–8 percentage points lower than that of judges below the median. This result is robust across a range of controls and specifications.

— INSERT TABLE 5 ABOUT HERE —

In parallel, the average cognitive uncertainty is substantially higher for high complexity applications than for low complexity applications. As previously mentioned, we expected the same directional evidence from the two measures of cognitive complexity, but not that they would have the same sized effects. When measuring complexity as disagreement in the score, high complexity applications exhibit an average cognitive uncertainty that is 12–13 percentage points higher than when judge disagreement is low (Table 6). When considering the text-based measure, the average cognitive uncertainty is 4–5 percentage points higher for high complexity application, than for low complexity applications. These results are robust across a range of controls and specifications.

— INSERT TABLE 6 ABOUT HERE —

Figure 6 summarizes these findings by plotting accuracy and precision — the inverse of the cognitive uncertainty — across score deciles, separately by expertise and complexity. All subgroups exhibit a clear value of information relative to the prior, confirming that judges' signals are informative. Across the three panels, the same ordering emerges: the value of information is consistently higher for more experienced judges and for less cognitively complex applications, whether complexity is measured by disagreement or textual complexity. The gain over the prior is largest at the extremes of the score distribution, particularly at the lower end, where signals most clearly separate low-quality ventures. Near the center of the distribution, where posterior beliefs are closest to the decision threshold, accuracy approaches the prior and residual cognitive uncertainty is highest. This pattern is consistent with the model's comparative statics: accuracy and precision are both governed by the signal-to-noise ratio, which expertise raises and complexity lowers, and the informational gain is greatest where signals are most decisive.

— INSERT FIGURE 6 ABOUT HERE —

7. Simulating Treatment Effects

We present a simulation exercise designed to assess how sensitive our results on the estimated Bayes classifier are to the assumption that the incubator has no treatment effect. We therefore explore counterfactual scenarios where classifying ventures as high quality is not only based on their ex-post economic performance but also on the treatment effect from being admitted to the Incubateur.

We generate counterfactual scenarios where non-admitted ventures receive a boost in their economic performance, as if they had been admitted to the incubation program and benefited from its treatment. We simulate counterfactual values for different treatment effect sizes on survival probability, FTE, and funding for the non-admitted firms, while preserving the original values for the admitted firms. The technical details can be found in Appendix J.

Increasing treatment effects sizes, especially on survival and funding outcomes, has an impact on individual ventures' relative ranking. However, changes to their classification as a high-quality firm are limited, ranging between a reclassification of between 3% to 7% of all ventures while it changes the classification for between 2% and 18% for the admitted ventures. Table L4 in the Appendix display the proportion of ventures whose classification as high-quality changes under different simulated standardized treatment effect sizes for survival, employment and raised capital. Values in parenthesis denote the fraction of changes occurring among admitted ventures.

Table L5 in the Appendix presents the changes of classification accuracy rates for the estimated Bayes as well as for the judges' recommendations across the various simulation scenarios, compared to the point estimate for the baseline scenario with no treatment effect. Notably, the highest model accuracy is achieved when we do not apply any treatment effect. Therefore, the accuracy estimates that are presented in the main body of the paper are likely upper bounds and should be interpreted with some caution. Finally, the variation in accuracy across different treatment effect scenarios is limited, suggesting that our results are relatively robust to changes in treatment effect sizes.

8. Discussion

We examine judges' ability to balance the selection of high-opportunity ventures with the rejection of low-opportunity ventures in an environment characterized by substantial cognitive complexity and limited feedback on decision quality. Consistent with prior evidence that predicting entrepreneurial success is inherently difficult (McKenzie and Sansone, 2019), judges achieve only moderate accuracy, exceeding random classification by at most 10–12 percentage points. Notably, this level of performance closely matches that of an optimal Bayesian classifier trained on the same information, suggesting that limited predictive accuracy primarily reflects the low informativeness of signals contained in the applications.

Modeling judges as Bayesian decision-makers provides a useful organizing framework. A simple

threshold rule applied to the sum of a multidimensional score vector reproduces approximately 79% of judges' recommendations, increasing to 85% when allowing for modest asymmetry in misclassification costs. Residual discrepancies are concentrated in intermediate score ranges, where posterior uncertainty is highest. These patterns indicate that judges broadly adhere to Bayesian logic while operating near the informational frontier imposed by noisy signals.

We document heterogeneity in decision quality along the two dimensions emphasized by the model: evaluator expertise and venture complexity. We proxy expertise by judge fixed effects and by a judge's cumulative number of evaluations, decomposed into within- and between-judge components. Within judge, additional experience does not improve accuracy. Between judges, those who evaluate more ventures are more accurate on average. Because judges receive no systematic feedback on venture outcomes, we read this contrast as evidence of selection—more able or motivated evaluators conduct more evaluations—rather than learning-by-doing.

Cognitive complexity plays a more substantial role. Prediction accuracy declines sharply with increasing complexity, whether measured by cross-judge disagreement or textual characteristics of applications, and decision uncertainty rises correspondingly. Moreover, the returns to expertise are greatest at intermediate levels of complexity, consistent with theoretical predictions that expertise is most valuable when problems are neither trivial nor intractable.

Taken together, the results suggest that judges perform reasonably well given the informational constraints they face. Consequently, improvements in decision quality are unlikely to arise primarily from changes in decision rules or increased experience alone, but rather from enhancements in the quality and structure of available signals. While training and feedback may yield incremental gains, the proximity of observed behavior to the Bayesian benchmark implies that substantial improvements depend on improving signal extraction.

These findings underscore the importance of designing effective information environments. This includes eliciting relevant information from applicants, structuring evaluation criteria,

calibrating scoring systems, and providing decision-makers with tools that facilitate signal extraction while minimizing unnecessary cognitive complexity.

One implication is that increasing the number of independent evaluations per application can mitigate the effects of cognitive noise. As shown by (Kaplan et al., 2008), larger sample sizes improve the precision of aggregated scores by averaging out idiosyncratic variation across judges, even though they do not reduce the underlying noise in individual assessments.

Another approach is to strengthen incentives for accurate evaluation. For example, recent work explores alternative incentive mechanisms in committee-based selection processes (Singh Chawla, 2023). Mechanisms such as "golden tickets," which allow individual evaluators to override majority decisions, are designed to mitigate Type II errors by preserving high-potential candidates that might otherwise be rejected. Similar mechanisms could be adapted to this setting by allocating a limited number of veto rights to judges.[10]

We do not observe, nor do we attempt to recover, the judges' underlying decision rules. However, alternative decision models or heuristics beyond Bayesian updating may be identifiable in the data. Moreover, judges may not fully utilize all information contained in the applications; some signals could remain unexploited yet be extractable using methods such as natural language processing or large language models. We leave the exploration of these possibilities to future research.

Finally, our analysis focuses on the identification of high-performing ventures, which may not align with the objectives of public policy. From a policy perspective, the relevant criterion may instead be the marginal impact of support on venture outcomes (McKenzie and Sansone, 2019). This distinction is reflected in the multi-stage structure of the incubator process: the first stage, which we analyze, screens for expected performance, while subsequent stages refine selection based on potential impact and portfolio considerations.

[10]Committee voting do apply in stage 2 and 3 of the HEC Incubateur process but those stages are left unanalyzed in this article. This omission does not in any way impact results or inferences made on the first stage process.

References

Anderson, J., 1983. Lix and rix: Variations on a little-known readability index. Journal of Reading 26, 490–496.

Aragones, E., Gilboa, I., Postlewaite, A., Schmeidler, D., 2005. Fact-free learning. American Economic Review 95, 1355–1368.

Arrieta, G., Nielsen, K., 2024. Procedural decision-making in the face of complexity. Technical Report. Working Paper.

Assenova, V.A., Amit, R., 2024. Poised for growth: Exploring the relationship between accelerator program design and startup performance. Strategic Management Journal .

Åstebro, T., Elhedhli, S., 2006. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. Management Science 52, 395–409.

Åstebro, T., Koehler, D.J., 2007. Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. Journal of Behavioral Decision Making 20, 381–403.

Augenblick, N., Lazarus, E., Thaler, M., 2025. Overinference from weak signals and underinference from strong signals. The Quarterly Journal of Economics 140, 335–401.

Avnimelech, G., Dushnitsky, G., Ellsaesser, F., Fitza, M., 2025. Are accelerators akin to breweries or wineries? a bayesian variance decomposition of accelerator and cohort effects. Strategic Management Journal 46, 534–579.

Banovetz, J., Oprea, R., 2023. Complexity and procedural choice. American Economic Journal: Microeconomics 15, 384–413.

Benjamin, D.J., 2019. Chapter 2 - errors in probabilistic reasoning and judgment biases, in: Bernheim, B.D., DellaVigna, S., Laibson, D. (Eds.), Handbook of Behavioral Economics - Foundations and Applications 2. North-Holland. volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pp. 69–186.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer.

Bernheim, D., Lucia, A., Nielsen, K., Sprenger, C.D., 2026. When Are Decisions Improvable? An Evaluation of Diagnostic Methods. Technical Report. National Bureau of Economic Research.

Blackwell, D., 1953. Equivalent comparisons of experiments. The annals of mathematical statistics , 265–272.

Börgers, T., Hernando-Veciana, A., Krähmer, D., 2013. When are signals complements or substitutes? Journal of Economic Theory 148, 165–195.

Butler, D.J., Loomes, G.C., 2007. Imprecision as an account of the preference reversal phenomenon. American Economic Review 97, 277–297. URL: https://www.aeaweb.org/articles?id=10.1257/aer.97.1.277, doi:10.1257/aer.97.1.277.

Chavda, A., Gans, J.S., Stern, S., 2024. Theory-based entrepreneurial search. Strategy Science 9, 397–415.

Chu, J.Y., Voelkel, J.G., Stagnaro, M.N., Kang, S., Druckman, J.N., Rand, D.G., Willer, R., 2024. Academics are more specific, and practitioners more sensitive, in forecasting interventions to strengthen democratic attitudes. Proceedings of the National Academy of Sciences 121, e2307008121.

de Clippel, G., Moscariello, P., Ortoleva, P., Rozen, K., 2024. Caution in the Face of Complexity. Technical Report. Mimeo.

Cohen, S., Fehder, D.C., Hochberg, Y.V., Murray, F., 2019a. The design of startup accelerators. Research Policy 48, 1781–1797.

Cohen, S., Hochberg, Y.V., 2014. Accelerating startups: The seed accelerator phenomenon. SSRN Working Paper .

Cohen, S., Koning, R., 2024. Advice and the bayesian entrepreneur .

Cohen, S.L., Bingham, C.B., Hallen, B.L., 2019b. The role of accelerator designs in mitigating bounded rationality in new ventures. Administrative Science Quarterly 64, 810–854.

Criscuolo, P., Dahlander, L., Grohsjean, T., Salter, A., 2017. Evaluating novelty: The role of panels in the selection of r&d projects. Academy of Management Journal 60, 433–460.

Dahlin, K.B., Chuang, Y.T., Roulet, T.J., 2018. Opportunity, motivation, and ability to learn from failures and errors: Review, synthesis, and ways to move forward. Academy of management annals 12, 252–277.

Dawes, R.M., 1975. Graduate admission variables and future success: We cannot tell whether the standard selection measures used by graduate schools are valid. Science 187, 721–723.

Dawes, R.M., 1979. The robust beauty of improper linear models in decision making. American psychologist 34, 571.

Dawes, R.M., Corrigan, B., 1974. Linear models in decision making. Psychological bulletin 81, 95.

Dawes, R.M., Faust, D., Meehl, P.E., 1989. Clinical versus actuarial judgment. Science 243, 1668–1674.

DellaVigna, S., Pope, D., 2018. Predicting experimental results: who knows what? Journal of Political Economy 126, 2410–2456.

Devroye, L., Györfi, L., Lugosi, G., 2013. A probabilistic theory of pattern recognition. volume 31. Springer Science & Business Media.

Dobbie, W., Song, J., 2015. Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. American economic review 105, 1272–1311.

Drerup, T., Enke, B., von Gaudecker, H.M., 2017. The precision of subjective data and the explanatory power of economic models. Journal of Econometrics 200, 378–389. URL: https://www.sciencedirect.com/science/article/pii/S0304407617301033, doi:https://doi.org/10.1016/j.jeconom.2017.06.017. measurement Error Models.

D'acunto, F., Hoang, D., Paloviita, M., Weber, M., 2023. Iq, expectations, and choice. The Review of Economic Studies 90, 2292–2325.

Enke, B., 2024. The cognitive turn in behavioral economics. Technical Report. Working Paper.

Enke, B., Graeber, T., 2023. Cognitive uncertainty. The Quarterly Journal of Economics 138, 2021–2067.

Fischhoff, B., 1975. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human perception and performance 1, 288.

Franzoni, C., Guerini, M., Nørding Christensen, R., Stanaj, A., 2025. Expert predictions and errors in research funding decisions. Academy of Management Proceedings 2025, 11186.

Giglio, S., Maggiori, M., Stroebel, J., Utkus, S., 2021. Five facts about beliefs and portfolios. American Economic Review 111, 1481–1522. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20200243, doi:10.1257/aer.20200243.

Gius, L., 2025. Disagreement predicts startup success: Evidence from venture competitions. Strategy Science 10, 93–108.

Gonzalez-Uribe, J., Leatherbee, M., 2018. The effects of business accelerators on venture performance: Evidence from start-up chile. The Review of Financial Studies 31, 1566–1603.

González-Uribe, J., Reyes, S., 2021. Identifying and boosting "gazelles": Evidence from business accelerators. Journal of Financial Economics 139, 260–287.

Grether, D.M., 1980. Bayes rule as a descriptive model: The representativeness heuristic. The Quarterly journal of economics 95, 537–557.

Grove, W.M., Meehl, P.E., 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. Psychology, public policy, and law 2, 293.

Halevy, Y., Mayraz, G., 2024. Identifying rule-based rationality. The Review of Economics and Statistics 106, 1369–1380. URL: https://doi.org/10.1162/rest_a_01232, doi:10.1162/rest_a_01232, arXiv:https://direct.mit.edu/rest/article-pdf/106/5/1369/2468957/rest_a_01232.pdf.

Hallen, B.L., Cohen, S.L., Bingham, C.B., 2020. Do accelerators work? if so, how? Organization Science 31, 378–414.

Howell, S.T., 2020. Reducing information frictions in venture capital: The role of new venture competitions. Journal of Financial Economics 136, 676–694.

Howell, S.T., 2021. Learning from feedback: Evidence from new ventures. Review of Finance 25, 595–627.

Jovanovic, B., Nyarko, Y., 1995. A bayesian learning model fitted to a variety of empirical learning curves. Brookings Papers on Economic Activity. Microeconomics 1995, 247–305.

Kahneman, D., 2018. Comment on "artificial intelligence and behavioral economics", in: Agrawal, A., Gans, J., Goldfarb, A. (Eds.), The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, Chicago, pp. 608–610.

Kahneman, D., Sibony, O., Sunstein, C.R., 2021. Noise: a flaw in human judgment. Little, Brown Spark., New York.

Kaplan, D., Lacetera, N., Kaplan, C., 2008. Sample size and precision in nih peer review. PLoS One 3, e2761.

Kendall, C., Oprea, R., 2024. On the complexity of forming mental models. Quantitative Economics 15, 175–211.

Kerr, W.R., Nanda, R., 2015. Financing innovation. Annual review of financial economics 7, 445–462.

Kling, J.R., 2006. Incarceration length, employment, and earnings. American Economic Review 96, 863–876.

Lane, J., Rietzler, N., 2026. DESIGNING EVALUATION PANELS: WHEN DOES PROFESSIONAL PANEL DIVERSITY IMPROVE STARTUP SELECTION? Technical Report.

Lane, J.N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E.C., Lakhani, K.R., 2022. Conservatism gets funded? a field experiment on the role of negative information in novel project evaluation. Management science 68, 4478–4495.

Lerner, J., Malmendier, U., 2013. With a little help from my (random) friends: Success and failure in post-business school entrepreneurship. The Review of Financial Studies 26, 2411–2452.

Maestas, N., Mullen, K.J., Strand, A., 2013. Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. American economic review 103, 1797–1829.

Mazur, J.E., Hastie, R., 1978. Learning as accumulation: a reexamination of the learning curve. Psychological bulletin 85, 1256.

McKenzie, D., Sansone, D., 2019. Predicting entrepreneurial success is hard: Evidence from a business plan competition in nigeria. Journal of Development Economics 141, 102369.

Molavi, P., Tahbaz-Salehi, A., Vedolin, A., 2023. Model complexity, expectations, and asset prices. The Review of Economic Studies 91, 2462–2507. URL: https://doi.org/10.1093/restud/rdad073, doi:10.1093/restud/rdad073, arXiv:https://academic.oup.com/restud/article-pdf/91/4/2462/58447653/rdad073.pdf.

Moore, D.A., Swift, S.A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H.H., Tenney, E.R., 2017. Confidence calibration in a multiyear geopolitical forecasting competition. Management Science 63, 3552–3565.

Murphy, A.H., 1973. A new vector partition of the probability score. Journal of Applied Meteorology and Climatology 12, 595–600.

Murphy, A.H., Winkler, R.L., 1984. Probability forecasting in meteorology. Journal of the American Statistical Association 79, 489–500.

Nanda, R., 2024. Priors, experiments, learning and persuasion in (bayesian) entrepreneurial finance .

Oprea, R., 2020. What makes a rule complex? American economic review 110, 3913–3951.

Oprea, R., 2024. Complexity and Its Measurement. Technical Report. UC Santa Barbara.

Ortoleva, P., 2022. Alternatives to bayesian updating. Annual Review of Economics 15, 1–28.

Sampat, B., Williams, H.L., 2019. How do patents affect follow-on innovation? evidence from the human genome. American Economic Review 109, 203–236.

Schwartz, S., Griffin, T., 2012. Medical thinking: The psychology of medical judgment and decision making. Springer Science & Business Media.

Scott, E.L., Shu, P., Lubynsky, R.M., 2020. Entrepreneurial uncertainty and expert evaluation: An empirical analysis. Management Science 66, 1278–1299.

Sharapov, D., Dahlander, L., 2025. Selection regimes and selection errors. Organization Science .

Singh Chawla, D., 2023. 'golden tickets' on the cards for nsf grant reviewers.

Smith, J.F., Kida, T., 1991. Heuristics and biases: Expertise and task realism in auditing. Psychological bulletin 109, 472.

Thompson, P., 2010. Learning by doing. Handbook of the Economics of Innovation 1, 429–476.

Tong, Y.L., 2012. The multivariate normal distribution. Springer Science & Business Media.

Waldman, J.D., Yourstone, S.A., Smith, H.L., 2003. Learning curves in health care. Health care management review 28, 41–54.

Yu, S., 2020. How do accelerators impact the performance of high-technology ventures? Management Science 66, 530–552.

Zacharakis, A.L., Meyer, G.D., 2000. The potential of actuarial decision models: can they improve the venture capital investment decision? Journal of Business venturing 15, 323–346.

Zacharakis, A.L., Shepherd, D.A., 2001. The nature of information and overconfidence on venture capitalists' decision making. Journal of business venturing 16, 311–332.

Tables

Panel A: General Statistics	
Number of companies	579
HEC graduate	40%
Age	33
Female	46%
Team size	4 (2)
Incorporated	64%
Panel B: Business Model	
Subscription	54%
Marketplace	21%
(e)Commerce	15%
Agency / Training center	4%
Media/Social networks	3%
Mobile App	3%
Panel C: Maturity	
Minimum Viable Product	46%
Prototype	28%
Full working Product	26%
Panel D: Sector	
Software IT	23.7%
Finance Real Estate	12.4%
LifeScience Healthcare	11.4%
Hospitality Education	9.7%
Consulting Professional Services	9.3%
Retail Wholesale Distribution	6.9%
Consumer Goods	6.4%
Media Entertainment Culture	6.4%
Other	13.8%

Table 1: Applying ventures
The table presents summary statistics of ventures that pass through the eligibility criteria at the HEC Incubateur. Panel A provides general statistics on the applicants, including the total number of applications, the proportion of HEC graduates among founders, the average founder age, and the likelihood of having at least one female co-founder. It also reports the average team size at the time of application, with the average number of co-founders shown in parentheses, and the percentage of ventures already legally registered as incorporated. Panel B displays the distribution of business models among applicants, ranking them by prevalence. Panel C categorizes the maturity stage of the applicants' products, from early-stage prototypes to fully developed offerings. Panel D provides the share of sectors present among applying ventures.

		Survival	FTE		€Mln Funding		
Batch	N Ventures	Mean	Mean	$P75$	$P > 0$	Mean$\mid > 0$	$P75 \mid > 0$
2021 Fall	62	56%	10.9	15.5	11%	13	6.2
2021 Summer	49	71%	10.3	12	12%	12.7	2.9
2022 Spring	60	82%	5.9	9	15%	0.7	1
2022 Summer	96	64%	8.2	9	12%	1.2	1.6
2022 Winter	66	74%	10.2	13	23%	2.3	1.7
2023 Spring	72	85%	4.8	6	4%	0.4	0.6
2023 Summer	66	76%	3.9	5	9%	0.7	1
2023 Winter	54	81%	3.5	4	7%	0.6	0.7
2024 Winter	54	72%	3.8	5	7%	1	1
All	579	73%	6.7	8	11%	3.5	2

Table 2: Ventures Survival Rates, FTE and Funding
The table presents summary statistics for all ventures evaluated across all batches in the dataset. Column (2) reports the number of ventures evaluated; column (3) shows the survival rate, i.e., the percentage of ventures in each batch still active as of Summer 2024. Columns (4)-(5) shows the average and the 75th percentile of FTEs among surviving ventures. Columns (6)-(8) provide metrics on post application funding in millions of euros: the probability of receiving any funding post-application, the mean funding amount if received, and the funding amount for the top 25% by post application funding in the batch.

Dependent Variable:	Accuracy				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
$Log(\hat{e}_{i,j}) - \overline{Log(\hat{e}_j)}$	-6.933**	-7.000**	-5.071*	-9.130**	-12.13***
	(2.752)	(2.740)	(2.993)	(3.550)	(3.859)
$\overline{Log(\hat{e}_j)}$	5.666*	5.624*	1.604	4.398	
	(3.159)	(3.177)	(3.871)	(3.052)	
Fixed-effects					
Batch (9)			Yes	Yes	Yes
Venture (579)				Yes	Yes
Evaluator (72)					Yes
Control: $\hat{s}_{i,j}$		Yes	Yes	Yes	Yes
Fit statistics					
Observations	1,638	1,638	1,638	1,638	1,638

Table 3: Accuracy and Judges' Experience
This table reports estimates from the regression:

$$\text{Accuracy}_{i,j} = \alpha + \beta_1 \left(\log(\hat{e}_{i,j}) - \overline{\log(\hat{e}_j)} \right) \tag{13}$$

$$+ \beta_2 \overline{\log(\hat{e}_j)} + \beta_3 \hat{s}_{i,j} + \theta_i + \delta_{t(i)} + \varepsilon_{i,j}, \tag{14}$$

where $\hat{e}_{i,j}$ denotes the cumulative number of evaluations previously performed by judge j at the time of evaluating venture i, $\hat{e}_j$ is the average number of evaluations made by judge j, $\hat{s}_{i,j}$ are evaluations' scores, $\delta_{t(i)}$ batch fixed effects and θ_i venture fixed effects. Regressions are estimated on the entire sample of evaluations with OLS. The variable $\log(\hat{e}_{i,j}) - \overline{\log(\hat{e}_j)}$ captures the within-judge component of experience and measures how a judge's prediction accuracy changes as the number of ventures evaluated by the same judge increases relative to that judge's own average experience. The variable $\overline{\log(\hat{e}_j)}$ captures the between-judge component and measures whether judges who evaluate more ventures on average are systematically more accurate than other judges. Standard errors are clustered at the venture and evaluator level. Coefficients are reported in percentage points. Observations are inversely weighted by the total number of evaluations performed by each judge in order to account for the concentration of evaluations among a small number of judges. Standard errors are clustered at the venture and evaluator level. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Dependent Variable:	Accuracy							
	Disagreement				Text-based			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Complexity$_i, Q2$	-3.185	-3.238	-1.954	-2.218	-7.585*	-6.505	-5.841	-5.806
	(3.049)	(3.038)	(2.837)	(3.123)	(4.278)	(4.137)	(4.043)	(4.173)
Complexity$_i, Q3$	-10.63***	-10.87***	-7.724**	-7.783***	-10.87***	-9.839**	-8.036**	-7.544**
	(3.129)	(3.153)	(2.989)	(2.908)	(3.827)	(3.720)	(3.544)	(3.419)
Complexity$_i, Q4$	-9.255***	-10.94***	-8.098**	-8.747***	-6.436	-5.539	-5.715	-4.687
	(3.391)	(3.457)	(3.242)	(3.254)	(4.121)	(3.969)	(3.827)	(3.883)
Fixed-effects								
Batch (9)			Yes	Yes			Yes	Yes
Evaluator (72)				Yes				Yes
Controls: $\hat{s}_{i,j}$		Yes	Yes	Yes		Yes	Yes	Yes
Fit statistics								
Accuracy Complexity$_i, Q1$	69.21%	69.21%	69.21%	69.21%	69.34%	69.34%	69.34%	69.34%
Observations	1,638	1,638	1,638	1,638	1,638	1,638	1,638	1,638

Table 4: Accuracy and Venture Complexity
This table reports estimates from the regression:

$$\text{Accuracy}_{i,j} = \alpha + \sum_{f=2}^{4} \beta_f \hat{\sigma}_i^f + \hat{s}_{i,j} + \delta_{t(i)} + \gamma_j + \varepsilon_{i,j}$$

where $\hat{\sigma}_i^f$ are indicators for quartiles of venture complexity, $\hat{s}_{i,j}$ are evaluations' scores, $\delta_{t(i)}$ batch fixed effects and γ_j are judge fixed effects. The reference category is the 1st quartile, corresponding to ventures with the lowest level of complexity. Regressions are estimated on the entire sample of evaluations. Columns (1)–(4) report estimates from OLS regressions using the disagreement based complexity measure; columns (5)-(8) use the text based complexity measure RIX. Coefficients represent percentage point differences in accuracy relative to the reference group. Standard errors (in parentheses) are clustered at the judge and venture level. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Dependent Variable:	Uncertainty			
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	0.2360***	0.2349***	0.2417***	
	(0.0012)	(0.0008)	(0.0007)	
High Experience$_j$	-0.0192***	-0.0169***	-0.0170***	-0.0174***
	(0.0027)	(0.0010)	(0.0007)	(0.0011)
Fixed-effects				
Batch (9)				Yes
Venture (579)				Yes
Controls: $\hat{s}_{i,j}$		Yes	Yes	Yes
Controls: $\hat{s}^2_{i,j}$			Yes	Yes
Fit statistics				
Observations	1,638	1,638	1,638	1,638

Table 5: Uncertainty and Judges Experience
This table reports estimates from the regression:

$$\text{Var}(\hat{v}_j \mid \hat{s}_{i,j}) = \alpha + \beta_1 \text{High Experience}_j + \beta_2 \hat{s}_{i,j} \tag{15}$$

$$+ \beta_3 \hat{s}^2_{i,j} + \delta_{t(i)} + \theta_i + \varepsilon_{i,j}. \tag{16}$$

where High Experience$_j$ equals 1 for judges above the median in the distribution of the number of cumulative evaluations, $\hat{e}_j$ is the average number of evaluations made by judge j, $\hat{s}_{i,j}$ are evaluations' scores, $\delta_{t(i)}$ batch fixed effects and θ_i venture fixed effects. Regressions are estimated on the entire sample of evaluations with OLS. Standard errors are clustered at the venture and evaluator level. Standard errors are clustered at the venture and evaluator level. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Dependent Variable:	Uncertainty							
	Disagreement				Text-based			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Constant	0.2120***	0.2101***	0.2182***		0.2211***	0.2220***	0.2331***	
	(0.0039)	(0.0020)	(0.0021)		(0.0017)	(0.0008)	(0.0008)	
High Complexity$_i$	0.0262***	0.0299***	0.0296***	0.0303***	0.0115***	0.0098***	0.0093***	0.0094***
	(0.0036)	(0.0027)	(0.0028)	(0.0030)	(0.0018)	(0.0010)	(0.0007)	(0.0009)
Fixed-effects								
Batch (9)				Yes				Yes
Evaluator (72)				Yes				Yes
Controls: $\hat{s}_{i,j}$		Yes	Yes	Yes		Yes	Yes	Yes
Controls: $\hat{s}^2_{i,j}$			Yes	Yes			Yes	Yes
Fit statistics								
Observations	1,638	1,638	1,638	1,638	1,638	1,638	1,638	1,638

Table 6: Uncertainty and Venture Complexity
This table reports estimates from the regression:

$$\begin{aligned} \text{Var}(\hat{v}_j \mid \hat{s}_{i,j}) \; =& \alpha + \beta_1 \text{High Complexity}_i && (17) \\ & + \beta_2 \hat{s}_{i,j} + \beta_3 \hat{s}^2_{i,j} + \delta_{t(i)} + \gamma_j + \varepsilon_{i,j}. && (18) \end{aligned}$$

where High Complexity$_i$ equals 1 for judges above the median in the distribution of the complexity measure, $\hat{s}_{i,j}$ are evaluations' scores, $\delta_{t(i)}$ batch fixed effects and γ_j judge fixed effects. Columns (1)–(4) report estimates from OLS regressions using the disagreement based complexity measure; columns (5)-(8) use the text based complexity measure RIX. Standard errors are clustered at the venture and evaluator level. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figures

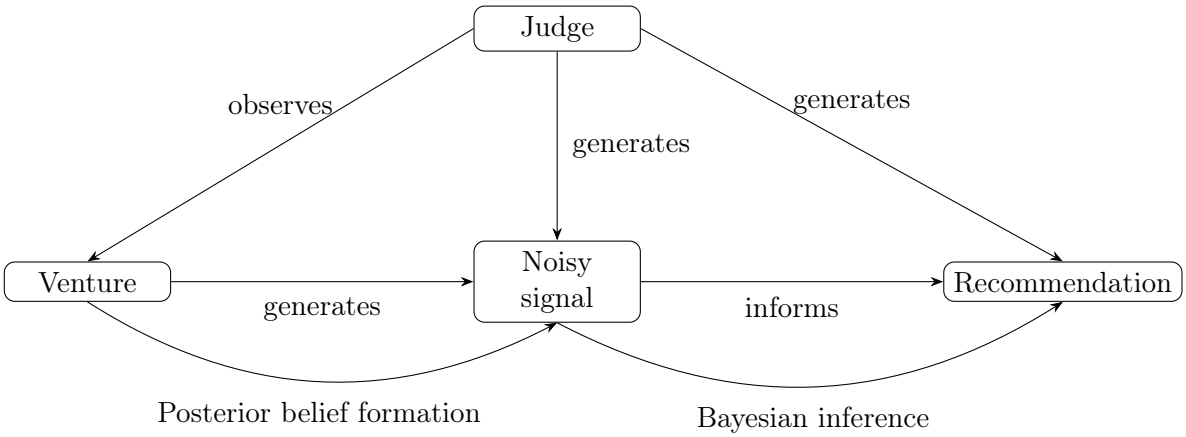


Figure 1: Conceptual model
Stylized representation of the judges' model of inference and their decision process.

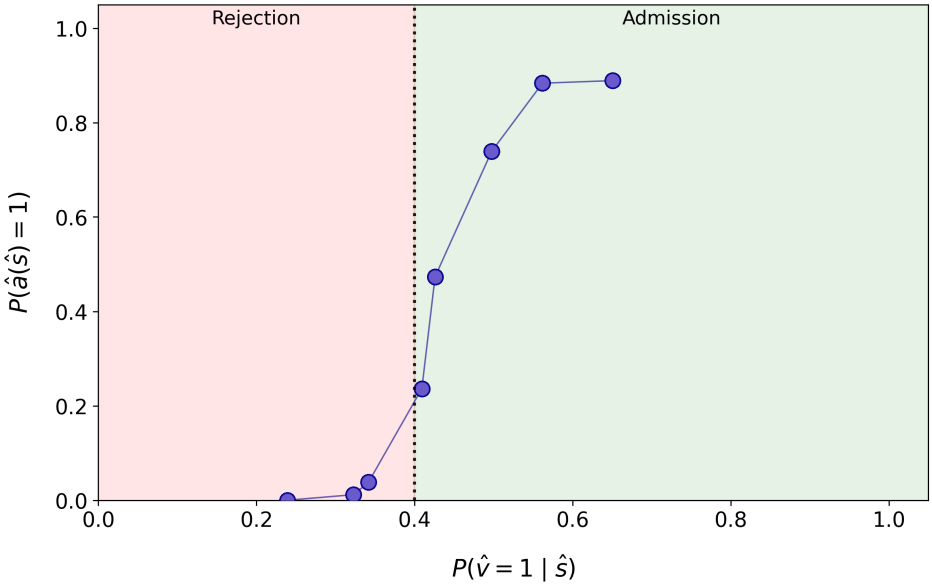


Figure 2: Recommendation rates and venture success.
The figure plots, across eight score bins, the empirical recommendation rate $P(\hat{a}(\hat{s}) = 1)$ against the corresponding ex-post probability of being top quality within the venture's batch, $P(\hat{v} = 1 \mid \hat{s})$. Each bin contains between 8% and 17% of all 1638 evaluations. The green region indicates ventures for which the probabilities of being among the top quality exceeds 40%, while the red region indicates ventures for which it is at most 40%.

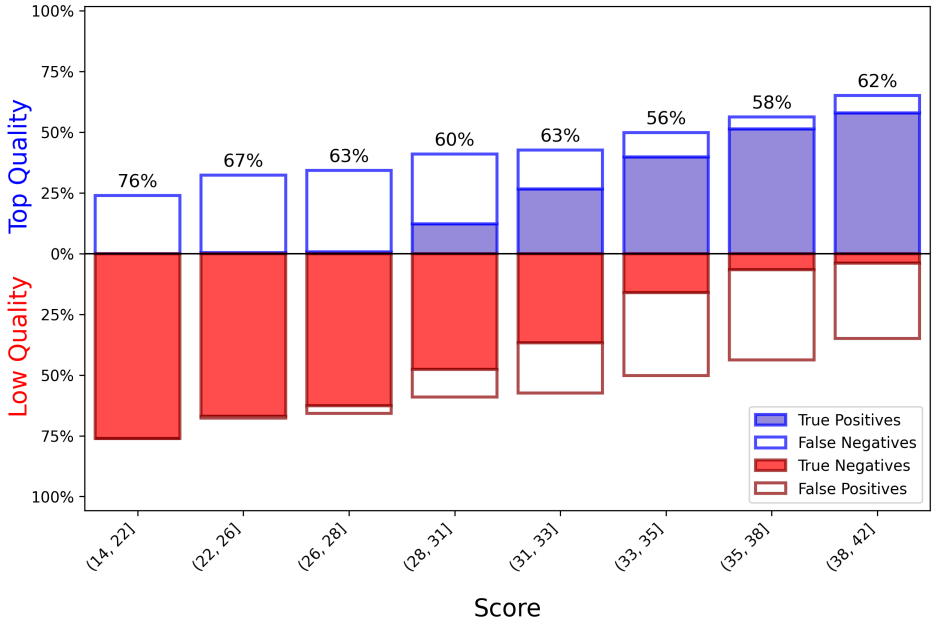


Figure 3: Decision Errors
The figure displays the ratio of evaluations that correctly or incorrectly classified a given venture as top quality in its batch, for different values of the score. We group the evaluations into eight bins according to their score. Each bin contains between 8% and 17% of all 1638 evaluations. For each bin we count the following: the ratio of top-quality ventures correctly identified (true positives) and missed (false negatives), as well as the ratio of low-quality ventures correctly identified (true negatives) and incorrectly classified as top quality (false positives) over the total number of evaluations. The labels above each bar report the accuracy rate within each bin.

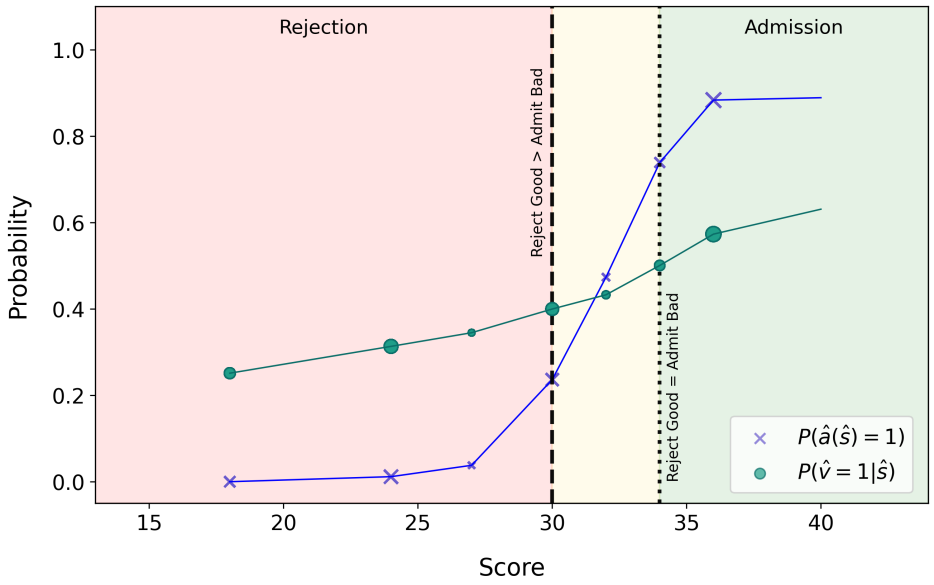


Figure 4: Score, Recommendation rates and Bayesian posterior
The figure plots (i) judges' recommendation rate $P(\hat{a}(\hat{s}) = 1)$ and (ii) the estimated posterior probability that a venture is top quality, $P(\hat{v} = 1 \mid \hat{s})$, as functions of the score. Both series are computed within eight score bins: crosses report the share of recommendations in each bin, and circles report the average estimated posterior within the same bin. Each bin contains between 11% and 14% of the 1638 evaluations. The shaded regions highlight Bayesian decision rules under different loss asymmetries: the green region corresponds to $\hat{\eta}(\hat{s}) \geq 0.5$ (Bayesian admission under symmetric costs), the red region to $\hat{\eta}(\hat{s}) \leq 0.4$ (Bayesian rejection under asymmetric costs), and the yellow region to $0.4 < \hat{\eta}(\hat{s}) < 0.5$, where admission is optimal only when the cost of rejecting a good venture exceeds the cost of admitting a bad one.

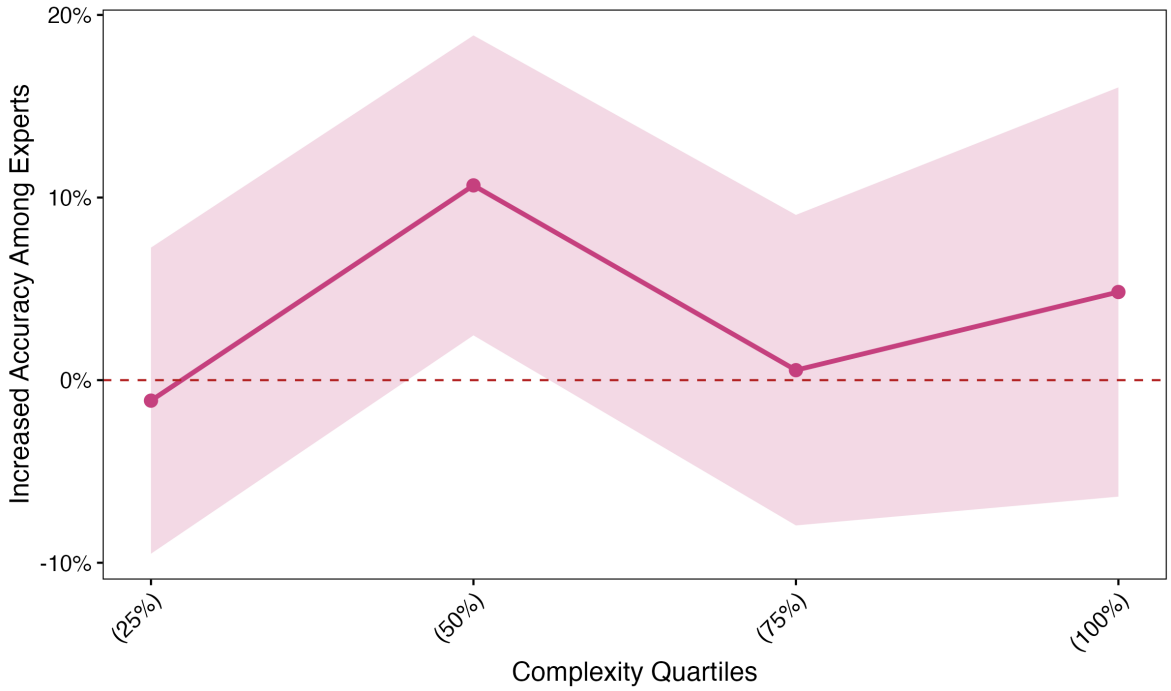


Figure 5: Accuracy Gain for Experienced Judges by Text Complexity
This figure plots point estimates and 95% confidence intervals for coefficient β_1 and $\beta_1 + \gamma_f$ for $\gamma = 2, 3, 4$ from the following linear regressions:

$$\text{Accuracy}_{i,j} = \alpha + \beta_1 \text{High Experience}_j + \sum_{f=2}^{4} \beta_f \hat{\sigma}_i^f + \sum_{f=2}^{4} \gamma_f \text{High Experience}_j \hat{\sigma}_i^f + \varepsilon_{i,j}$$

where $\hat{\sigma}_i^f$ are indicator variables for the 2nd, 3rd, and 4th quartiles of the RIX index of the application text complexity, and *High Experience* equals 1 for judges above the median of the distribution of the number of cumulative evaluations. The coefficients capture whether the accuracy gap between highly experienced and less experienced judges varies across levels of venture complexity. Each point estimate represents the percentage-point difference in accuracy between more and less experienced judges within a given complexity quartile. Confidence intervals are based on standard errors clustered at both the judge and venture levels.

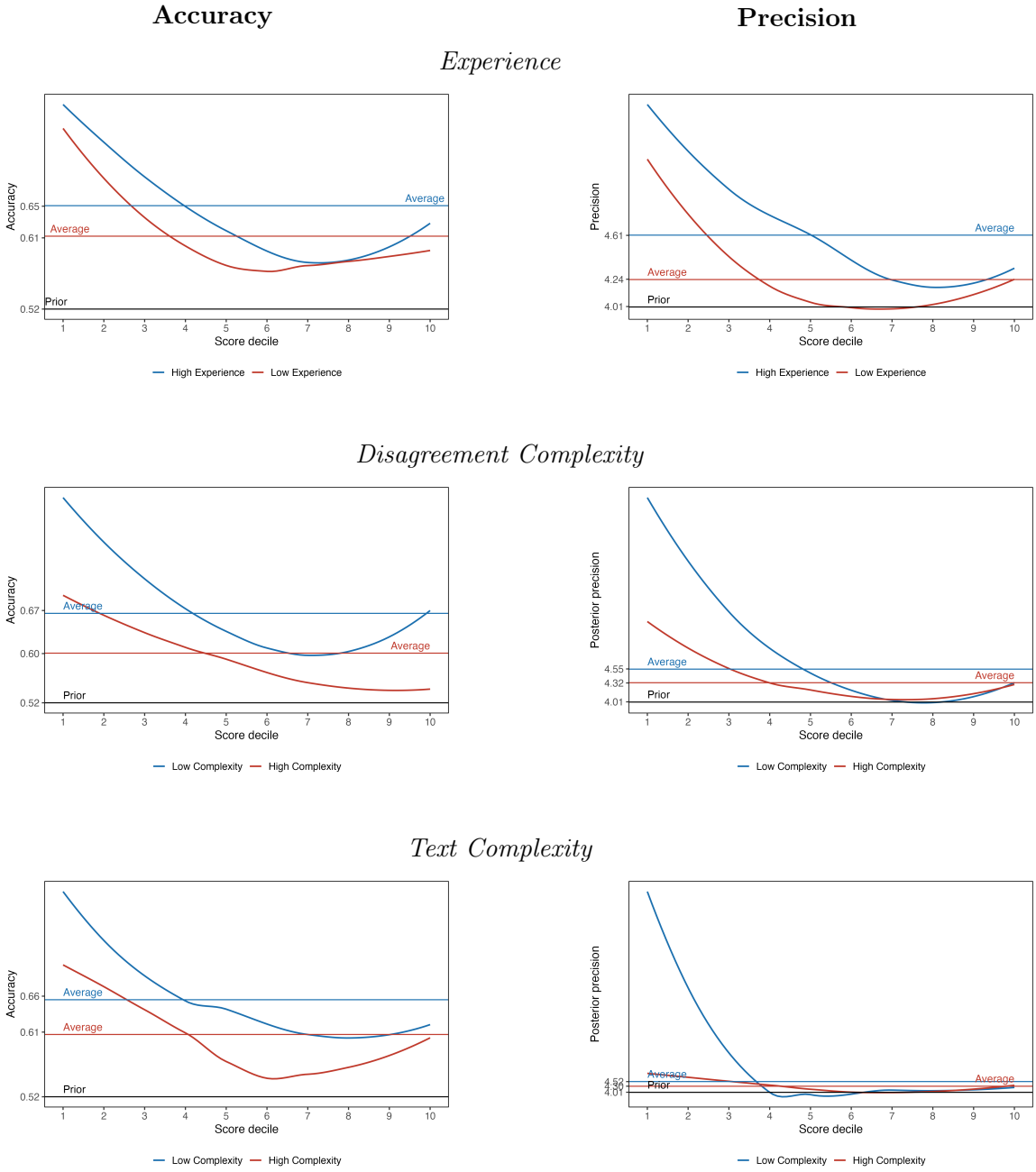


Figure 6: Value of Information, Experience, and Complexity
Each panel plots the relationship between the score and the information content of the corresponding signal. The left column reports the probability that the recommendation is correct (accuracy), while the right column reports precision, defined as the inverse of the cognitive uncertainty from the subgroup-specific empirical Bayes classifier. The top row compares evaluations produced by more versus less experienced judges. The middle row compares evaluations of ventures with high versus low disagreement-based complexity. The bottom row compares evaluations of ventures with high versus low text-based complexity, measured using the RIX index. In each panel, the horizontal black line indicates the prior benchmark, while the colored horizontal lines indicate subgroup averages. Blue lines correspond to lower-complexity ventures or more experienced judges, and red lines to higher-complexity ventures or less experienced judges.

Appendix

I. The Conceptual Model

The variable $w \sim \mathcal{N}(0,1)$ denotes the latent economic quality of a venture, while $v = \mathbf{1}_{w \geq \bar{w}}$ is a label that classifies a venture as good if its economic quality is high enough. Judges produce a noisy cognitive signal from the underlining economic quality: the error term is denoted by $\epsilon \sim \mathcal{N}\left(0, \frac{\sigma}{e}\right)$. A cognitive signal is modeled as

$$s = w + \epsilon.$$

Judges use a decision rule $a : \mathcal{S} \to [0,1]$, that to a signal realization associates a probability of recommending a venture, where $\mathcal{S} = \mathbb{R}$ is the set of possible signal realizations. It follows that signal distributes as $s \sim \mathcal{N}\left(0, 1 + \frac{\sigma}{e}\right)$ and the posterior distribution of the latent economic quality is

$$w|s \sim \mathcal{N}\left(\frac{e}{e+\sigma}s, \frac{\sigma}{e+\sigma}\right).$$

I.1. The decision problem

Consistently with the model formulation and the empirical setting, suppose that the incubator objective is to select the top companies as defined by the label $v = \mathbf{1}_{w \geq \bar{w}}$.

The decision problem is

$$\max_{a:\mathcal{S}\to[0,1]} \mathbb{E}_s[a(s)v + \gamma(1 - a(s))(1 - v)|s]$$

For analytical tractability, we intially assume $\bar{w} = 0$ and $\gamma = 1$, i.e. that the incubator aims at selecting the ventures that are better than average and give equal weight to misclassifications errors. Under more generic assumptions on $\bar{w}$ and γ, the derived close form solution is not valid anymore, however the results remain valid and testable by numerical computation.

Let $a^*(s)$ denote the optimal decision rule and

$$\xi(e,\sigma) = \mathbb{E}_s\left[a^*(s)v + \gamma(1-a^*(s))(1-v)\right]$$

the optimal accuracy as function of the signal distribution, $s \sim \mathcal{N}\left(0, 1+\frac{\sigma}{e}\right)$. Under the assumption that $\bar{w}=0$ and $\gamma=1$[11], the following holds:

Lemma 1. *The optimal decision rule is*

$$a^*(s) = \mathbf{1}_{s\geq 0}.$$

Proof.

From the classification problem with equal costs of misclassifications, one has:

$$a^*(s) = \begin{cases} 1 & \text{if } \mathbb{E}(v \mid s) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

[11] Under generic $\bar{w}$ and γ, the optimal decision rule is still a threshold rule, according to which a judge admits $(a(s)=1)$ whenever

$$s \geq \frac{e+\sigma}{e}\left[\bar{w} + \sqrt{\frac{\sigma}{e+\sigma}}\frac{\gamma}{1+\gamma}\right].$$

The threshold, that decreases in e and increases in σ, is informative about the optimal leniency of judges depending on expertise and complexity: more expert judges use a lower threshold (for a more informative score) than less expert, and more complex ventures imply a higher threshold than simpler ones.

Notice that

$$\mathbb{E}(v|s) \geq \frac{1}{2} \tag{19}$$

$$\iff \mathbb{P}(v = 1|s) \geq \frac{1}{2} \tag{20}$$

$$\iff \mathbb{P}(w \geq \bar{w}|s) \geq \frac{1}{2} \tag{21}$$

$$\iff \mathbb{P}\left(z \leq \frac{\mathbb{E}(w|s) - \bar{w}}{\sqrt{\text{Var}(w|s)}}\right) \geq \frac{1}{2} \quad z \sim \mathcal{N}(0,1) \tag{22}$$

$$\iff \frac{\mathbb{E}(w|s) - \bar{w}}{\sqrt{\text{Var}(w|s)}} \geq \Phi^{-1}\left(\frac{1}{2}\right) = 0 \tag{23}$$

$$\iff \mathbb{E}(w|s) \geq \bar{w} \tag{24}$$

$$\iff \frac{e}{e+\sigma} s \geq 0 \tag{25}$$

$$\iff s \geq 0. \tag{26}$$

□

In this setting, an explicit formula for the optimal expected accuracy can be derived.

Proposition 3. The optimal expected accuracy is

$$\xi(e, \sigma) = \frac{1}{2} + \frac{1}{\pi} \arctan \sqrt{\frac{e}{\sigma}}.$$

Proof. First, notice that

$$\begin{aligned}\xi(e,\sigma) &= \mathbb{E}_s\left[a^*(s) + (1-a^*(s))(1-v)\right] \\ &= \mathbb{E}_s\left[a^*(s)\left(2v-1\right) + (1-v)\right] \\ &= 2\mathbb{E}_s\left[a^*(s)v\right] - \mathbb{E}_s\left[a^*(s)\right] + \frac{1}{2} \\ &= 2\mathbb{E}_s\left[a^*(s)v\right] \\ &= 2\mathbb{P}\left(a^*(s)=1, v=1\right) \\ &= 2\mathbb{P}\left(s\geq 0, w\geq 0\right).\end{aligned} \tag{27}$$

Since (w,s) constitute a joint multivariate normal distribution with zero mean and correlation coefficient $\frac{1}{\sqrt{1+\frac{\sigma}{e}}}$, one has[12]

$$\begin{aligned}\mathbb{P}\left(s\geq 0, w\geq 0\right) &= \frac{1}{4} + \frac{1}{2\pi}\arcsin\frac{1}{\sqrt{1+\frac{\sigma}{e}}} \\ &= \frac{1}{4} + \frac{1}{2\pi}\arctan\sqrt{\frac{e}{\sigma}}.\end{aligned} \tag{28}$$

Hence, we conclude that

$$\xi(e,\sigma) = \frac{1}{2} + \frac{1}{\pi}\arctan\sqrt{\frac{e}{\sigma}}.$$

□

Since the arctan function is monotone, we conclude the following.

Proposition 4. $\xi(e,\sigma)$ is increasing in e and decreasing in σ.

The difference in accuracy between more and less experts, as a function of the

complexity of the underlining problem, can be studied by considering the derivative of the optimal accuracy,

$$\xi_e(e,\sigma) = \frac{1}{2\pi}\frac{1}{\sqrt{\sigma e}}\frac{\sigma}{e+\sigma}.$$

[12] See Tong (2012), Ch. 2, Page 14.

Such function represents the marginal value of expertise, that it is maximized at intermediate values of complexity and vanishes for very low and very high levels of complexity.

The cognitive uncertainty of an evaluator of a given expertise level evaluating a venture of a certain complexity is measured as

$$\mathbb{E}_s\left[\mathrm{Var}(v|s)\right].$$

Proposition 5. Cognitive uncertainty is increasing in σ and decreasing in e.

Proof. First, notice that

$$\mathbb{E}_s\left[\mathrm{Var}(v|s)\right] = \mathbb{E}_s\left[\mathbb{P}(v=1|s)\left(1-\mathbb{P}(v=1|s)\right)\right].$$

Moreover, notice that

$$\begin{aligned}\mathbb{P}[v=1|s] &= \mathbb{P}\left(w \geq 0 \mid s\right) \\ &= \mathbb{P}\left(z \leq s\frac{e}{\sqrt{\sigma(e+\sigma)}}\right) \quad z\sim\mathcal{N}(0,1) \\ &= \Phi\left(s\frac{e}{\sqrt{\sigma(e+\sigma)}}\right).\end{aligned} \tag{29}$$

We now apply a change of variable, noticing that

$$s \sim \sqrt{1+\frac{\sigma}{e}}z, \quad z\sim\mathcal{N}(0,1).$$

Therefore,

$$\begin{aligned}&\mathbb{E}_s\left[\mathrm{Var}(v|s)\right] \\ &= \mathbb{E}_s\left[\Phi\left(s\frac{e}{\sqrt{\sigma(e+\sigma)}}\right)\left(1-\Phi\left(s\frac{e}{\sqrt{\sigma(e+\sigma)}}\right)\right)\right] \\ &= \mathbb{E}_z\left[\Phi\left(z\sqrt{\frac{e}{\sigma}}\right)\left(1-\Phi\left(z\sqrt{\frac{e}{\sigma}}\right)\right)\right].\end{aligned} \tag{30}$$

Now, notice that the function $f(x) = x(1-x)$ is strictly concave and maximized at $x = \frac{1}{2}$,

furthermore $\Phi^{-1}(0) = \frac{1}{2}$. It follows that $\mathbb{E}_s[\text{Var}(v|s)]$ increases as $z\sqrt{\frac{e}{\sigma}}$ approaches zero, and it is therefore increasing in σ and decreasing in e. □

Remark. Proposition 4 and 5 can be equivalently proved by considering the informativeness of statistical experiments produced by judges' grading behavior (as argued in Appendix I). Consider a same judge evaluating two ventures with different venture specific noise. The judge has access to two signals

$$s = w + \epsilon, \quad s' = w + \epsilon'$$

where

$$\epsilon \sim \mathcal{N}\left(0, \frac{\sigma}{e}\right), \quad \epsilon \sim \mathcal{N}\left(0, \frac{\sigma'}{e}\right), \quad \sigma' > \sigma.$$

Such signals are produced by a judge evaluating two different ventures with two different complexity levels. One can notice that

$$s' = s + \nu, \quad \nu \sim \mathcal{N}\left(0, \frac{\sigma'}{e} - \frac{\sigma}{e}\right),$$

hence there exists a garbling between signal s and s' given by

$$s'|s \sim \mathcal{N}\left(s, \frac{\sigma'}{e} - \frac{\sigma}{e}\right).$$

Equivalently, for a fixed σ, two levels of expertise e and e' with $e > e'$ imply that signal s is more informative than signal s'. by Blackwell (1953), a more informative signal induces a higher value function for any decision problem, hence resulting in Proposition 4 when the decision problem concerns the maximization of the expected accuracy and in Proposition 5 when the decision problem concerns the minimization of the mean square error.

Remark. The model described above, while allowing for clean and tractable results, is restrictive. In particular, it assumes that the incubator only cares about whether ventures exceed a given quality threshold, without exhibiting any ordinal preference over how high that quality is. This modeling choice is motivated by consistency with the empirical treatment. Never-

theless, the qualitative insights extend beyond this specific setup. Indeed, suppose that the incubator cared about the cardinal value of economic quality. A plausible decision problem would be

$$\max_{a:\mathcal{S}\to[0,1]} \mathbb{E}_s \left[a(s)w + \frac{\bar{w}}{1-\bar{w}}(1-a(s))(1-w) \right].$$

Let $a^*(s)$ denote the optimal decision rule. It holds that $a^*(s) = 1$ if and only if $\mathbb{E}[w|s] \geq \bar{w}$. Hence, again incubator cares about selecting the top percentile of ventures that have economic quality higher than a threshold. Hence, under the assumption that $\bar{w} = 0$, the optimal decision rule is again $a^*(s) = \mathbf{1}_{s\geq 0}$. The optimal accuracy of the decision problem is

$$\begin{aligned}
\mathbb{E}_s \left[a^*(s)w \right] &= \mathbb{E}_s \left[\mathbf{1}_{s\geq 0} w \right] \\
&= \mathbb{E}_s \left[s\frac{e}{e+\sigma} \mid s \geq 0 \right] \\
&= \frac{e}{e+\sigma} \mathbb{E}_s[s \mid s \geq 0] \\
&= \sqrt{\frac{e}{e+\sigma}} \frac{\phi(0)}{1-\Phi(0)} \\
&= \sqrt{\frac{e}{e+\sigma}} \sqrt{\frac{2}{\pi}}.
\end{aligned} \tag{31}$$

where the stream of equalities comes from the expectation of a truncated normal distribution. Unsurprisingly, the expected optimal accuracy is again increasing in experience and decreasing in complexity. The marginal value of expertise – the derivative of the optimal expected accuracy with respect to the expertise level, is

$$\frac{1}{\sqrt{2\pi}} \sqrt{\frac{\sigma}{e}} \frac{\sqrt{\sigma}}{(e+\sigma)^{\frac{3}{2}}}.$$

Cognitive noise, defined as the expected posterior variance, is

$$\frac{\sigma}{\sigma+e}.$$

It increasing in σ, the complexity of a venture, and decreasing in e, the expertise of an evaluator.

Remark. Under equal misclassification costs, the optimal classification accuracy is related to the posterior variance through the simple first order approximation

$$\mathbb{E}_s\left[u(a^*(s), v)\right] \approx 1 - \mathbb{E}_s\left[\mathrm{Var}(v \mid s)\right]. \tag{32}$$

Conditioning on a signal s, the optimal expected accuracy is

$$u(a^*(s), v) = \max\left\{\mathbb{P}(v = 1 \mid s),\, 1 - \mathbb{P}(v = 1 \mid s)\right\}.$$

The expected optimal accuracy is related to the posterior variance through the following expression, whose first order approximation yields 32.

Claim 1. *The expected optimal accuracy can be expressed as*

$$\mathbb{E}_s\left[u(a^*(s), v)\right] = \frac{1}{2} + \mathbb{E}_s\left[\sqrt{\frac{1}{4} - \mathrm{Var}(v \mid s)}\right].$$

Proof. Let $x = \mathbb{P}(v = 1 \mid s)$. Then

$$u(a^*(s), v) = \max\{x, 1 - x\} = \frac{1}{2} + \left|x - \frac{1}{2}\right|.$$

The posterior variance is

$$\mathrm{Var}(v \mid s) = x(1 - x).$$

Rewriting,

$$x(1 - x) = \frac{1}{4} - \left(x - \frac{1}{2}\right)^2,$$

which implies

$$\left|x - \frac{1}{2}\right| = \sqrt{\frac{1}{4} - \mathrm{Var}(v \mid s)}.$$

Substituting into the expression for $u(a^*(s), v)$ yields

$$u(a^*(s), v) = \frac{1}{2} + \sqrt{\frac{1}{4} - \mathrm{Var}(v \mid s)}.$$

Taking expectations over s gives the result. □

I.2. Discussion on the model assumptions

Up to this stage, we remain agnostic about what specifically determines expertise and complexity, and instead focus on their consequences. In particular, expertise is defined by a greater ability to draw precise inferences, regardless of the type of project being evaluated. Conversely, complexity reduces the precision with which information can be extracted, for any type of evaluator. In particular, expertise may arise from repeated exposure to a specific task, measured by the cumulative number of evaluations performed and the cumulative level of complexity encountered. At the same time, expertise may also derive from related experience outside the focal task, such as one's profession, status, or broader domain of specialization. Alternatively, expertise may reflect incentives and stakes in achieving the correct resolution of the task, or inherent talent. In reality, expertise is likely to be vertical: heterogeneous levels of expertise may emerge depending on the specific task. For example, a professional physician may be better suited to evaluating biotech ventures. However, at this stage we abstract from such task-specific heterogeneity, given the particular pool of applicants and startups we study. These considerations are important and deserve future research.

Complexity, on the other hand, may derive from several sources. Our main proxy for complexity is disagreement, which in the specific model with additive normal noise is an asymptotically unbiased and consistent estimator. In parallel, textual complexity and technological novelty, measured by the venture's technological maturity, also contribute to complexity. All these dimensions interact in determining the precision of the cognitive signal an evaluator can extract, and consequently affect evaluation accuracy and cognitive noise.

The marginal value of expertise is relevant for at least two reasons. First, since expertise is always valuable, assigning tasks at their "sweet spot" allows one to screen evaluators who are more expert than others. Second, from an allocative efficiency perspective, identifying the sweet spot can inform how to assign tasks across evaluators when expert capacity is limited.

I.3. General Informational Environment

The model predictions extend beyond the Normal-Normal framework considered above. This section develops a more general theory of expertise and complexity based on assumptions about the informativeness of statistical experiments and the complementarity structure between expertise and project complexity.

Let $\mathcal{E} \times \mathcal{C} \subseteq \mathbb{R}^2$ be a compact convex set, where $e \in \mathcal{E}$ denotes an agent's expertise level and $\sigma \in \mathcal{C}$ denotes project complexity. Let $\mathcal{V}$ be a finite state space with prior $p \in \Delta(\mathcal{V})$, and let $\mathcal{S}$ be a signal space. A *statistical experiment* indexed by expertise and complexity is a family of signal distributions

$$q : \mathcal{E} \times \mathcal{C} \times \mathcal{V} \to \Delta(\mathcal{S}).$$

We write $q_{e,\sigma}$ for the experiment at expertise level e and complexity σ, and let $\succeq_B$ denote the Blackwell order on statistical experiments (Blackwell, 1953). Recall that $q_{e,\sigma} \succeq_B q_{e',\sigma'}$ if and only if there exists a Markov kernel $M : \mathcal{S} \to \Delta(\mathcal{S})$ such that $q_{e',\sigma'}(\cdot \mid v) = M\, q_{e,\sigma}(\cdot \mid v)$ for all $v \in \mathcal{V}$. A more Blackwell-informative experiment leads to weakly higher expected utility in every decision problem.

A decision rule is a measurable map $a : \mathcal{S} \to [0,1]$, and $u : [0,1] \times \mathcal{V} \to \mathbb{R}$ is a utility function. The *value function* $\xi : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$ denotes the expected utility attainable under an optimal decision rule,

$$\xi(e,\sigma) \;=\; \mathbb{E}_{v\sim p,\; s\sim q_{e,\sigma}(\cdot|v)}\left[\max_{a:\mathcal{S}\to[0,1]} u\big(a(s),\, v\big)\right].$$

We assume throughout that $\xi : \mathcal{E}\times\mathcal{C} \to \mathbb{R}$ is twice continuously differentiable in $(e,\sigma) \in \mathcal{E}\times\mathcal{C}$.

We say that expertise and complexity are *complements* at $(e,\sigma) \in \mathcal{E} \times \mathcal{C}$ if

$$\frac{\partial^2 \xi}{\partial e\, \partial \sigma}(e,\sigma) \;>\; 0,$$

and *substitutes* if the reverse inequality holds strictly. Economically, complementarity means that the marginal return to expertise rises with project complexity. Substitutability captures

the opposite case, where expertise and complexity provide diminishing joint returns.

A characterization of complementarity in terms of statistical experiments is derived from Börgers et al. (2013). Given expertise levels $\bar{e} > \underline{e}$ and complexity levels $\bar{\sigma} > \underline{\sigma}$, define two auxiliary experiments:

- s_S: reveals signal from $q_{\bar{e},\bar{\sigma}}$ with probability $\frac{1}{2}$ and from $q_{\underline{e},\underline{\sigma}}$ with probability $\frac{1}{2}$;
- s_C: reveals signal from $q_{\underline{e},\bar{\sigma}}$ with probability $\frac{1}{2}$ and from $q_{\bar{e},\underline{\sigma}}$ with probability $\frac{1}{2}$.

Complementarity across the four experiment pairs holds for every decision problem if and only if $s_C \succeq_B s_S$; substitutability holds if and only if $s_S \succeq_B s_C$.

The analysis rests on the following conditions.

Assumption 1 (Informativeness of expertise and simplicity)**.**

1. *Simpler projects are more informative:* for all $e \in \mathcal{E}$,

$$\bar{\sigma} \geq \underline{\sigma} \implies q_{e,\underline{\sigma}} \succeq_B q_{e,\bar{\sigma}}.$$

2. *Higher expertise is more informative:* for all $\sigma \in \mathcal{C}$,

$$\bar{e} \geq \underline{e} \implies q_{\bar{e},\sigma} \succeq_B q_{\underline{e},\sigma}.$$

3. *Complementarity–substitutability:* there exists a function $f : \mathcal{E} \to \mathcal{C}$ such that

$$\frac{\partial^2 \xi}{\partial e \, \partial \sigma}(e,\sigma) \begin{cases} > 0 & \text{if } \sigma < f(e), \\ = 0 & \text{if } \sigma = f(e), \\ < 0 & \text{if } \sigma > f(e). \end{cases}$$

Conditions 1 and 2 are natural monotonicity requirements: more expertise and lower complexity each generate more informative signals in the Blackwell sense. Condition 3 imposes

a regularity structure on the interaction between expertise and complexity: for each level of expertise e, the threshold $\sigma^*(e) \equiv f(e)$ separates a complementarity region (low complexity) from a substitutability region (high complexity). Beyond $\sigma^*(e)$, further increases in expertise can no longer compensate for the informational losses induced by higher complexity.

Remark. In the Normal-Normal model of 4, the function $f : \mathcal{E} \to \mathcal{C}$ is the identity function $f(e) = e$.

Proposition 6. Under Assumption 1, the following hold for all $(e, \sigma) \in \mathcal{E} \times \mathcal{C}$:

1. *Monotonicity in expertise:* $\bar{e} > \underline{e}$ implies $\xi(\bar{e}, \sigma) \geq \xi(\underline{e}, \sigma)$.
2. *Monotonicity in complexity:* $\bar{\sigma} > \underline{\sigma}$ implies $\xi(e, \underline{\sigma}) \geq \xi(e, \bar{\sigma})$.
3. *Interior maximizer of the marginal return to expertise:* for each $e \in \mathcal{E}$, the function $\sigma \mapsto \xi_e(e, \sigma)$ attains a unique interior maximum at $\sigma^*(e) = f(e)$.

Proof. Points 1 and 2. Both statements follow directly from the Blackwell theorem (Blackwell, 1953): since a more Blackwell-informative experiment yields weakly higher expected utility in every decision problem, conditions 1 and 2 of Assumption 1 imply that ξ is weakly increasing in e and weakly decreasing in σ.

Point 3. Fix $e \in \mathcal{E}$. By condition 3 of Assumption 1, $\xi_{e\sigma}(e, \sigma) > 0$ for $\sigma < f(e)$ and $\xi_{e\sigma}(e, \sigma) < 0$ for $\sigma > f(e)$. Hence $\sigma \mapsto \xi_e(e, \sigma)$ is strictly increasing on $[\cdot, f(e))$ and strictly decreasing on $(f(e), \cdot]$, so it attains its unique global maximum at $\sigma^*(e) = f(e)$. □

Proposition 6 delivers several insights. First, the value of a project is unambiguously increasing in the agent's expertise and decreasing in complexity: a consequence of the Blackwell ordering alone, independent of the specific decision problem. Second, and more subtly, the marginal return to expertise is maximized at an intermediate complexity level $\sigma^*(e)$. For projects simpler than $\sigma^*(e)$, expertise and complexity are complements: additional complexity raises the value of being expert. For projects more complex than $\sigma^*(e)$, they become substitutes:

complexity is so high that even greater expertise fails to recover the informational losses it induces.

The Normal-Normal model is a special case of this framework. The results on posterior variance minimization are recovered by setting $u(a, v) = -(a - v)^2$, which yields the mean-squared-error objective. Results on prediction accuracy under arbitrary error weights and priors follow analogously.

J. Treatment Effect Simulation

We simulate counterfactual values for different treatment effect sizes on outcomes Y: survival probability SP, employment L, and funding K for the non-admitted firms, while preserving the original values for the admitted firms. Specifically, we define the counterfactual outcomes as follows:

$$Y_i = \begin{cases} Y_i, & \text{if } i \text{ was admitted} \\ Y_{couterfactual,i} & \text{if } i \text{ was not admitted.} \end{cases} \tag{33}$$

For funding and full-time employees (FTE), we assume that the counterfactual values are equal to the observed ex-post outcomes plus an outcome-specific treatment effect,

$$Y_{couterfactual,i} = Y_i + \delta_Y. \tag{34}$$

For survival, we maintain the historical survival status for firms that survived and sample the counterfactual outcome from a Bernoulli distribution for firms that did not survive,

$$SP_{\text{counterfactual},i} = \begin{cases} SP_i, & \text{if } SP_i = 1 \\ \sim \text{Ber}\left(P_{\text{c}}\right) & \text{if } SP_i = 0, \end{cases} \tag{35}$$

where P_{c} is such that

$$\bar{SP}_{\text{counterfactual}} = P + \delta_Y,$$

with P representing the average survival rate among non-admitted firms and $\bar{SP}$ the the average survival rate in the counterfactual scenario.

The counterfactual specifications assume that treatment effects are homogeneous across firms and additive in nature. Scenarios vary depending on the assumed treatment effect size for each outcome. We consider outcome-specific standardized treatment effects γ_Y corresponding to

0%, 10%, 20%, and 30% of a standard deviation (SD) increase, consistent with the results on incubateurs' treatment effects found in (Gonzalez-Uribe and Leatherbee, 2018), and generate absolute treatment effects by multiplying them with the standard deviation for each outcome, after winsorizing the top 5% of values,

$$\delta_Y^s = \gamma_Y^s \cdot \sigma_Y. \tag{36}$$

As a result, we generate 64[13] counterfactual scenarios. Finally, we define a new labeling indicating whether a firm is in the top x% on the combined ranking of each counterfactual scenario.

[13]Derived by considering different combinations of four possible treatment effect sizes on the three outcome variables.

K. Additional Figures

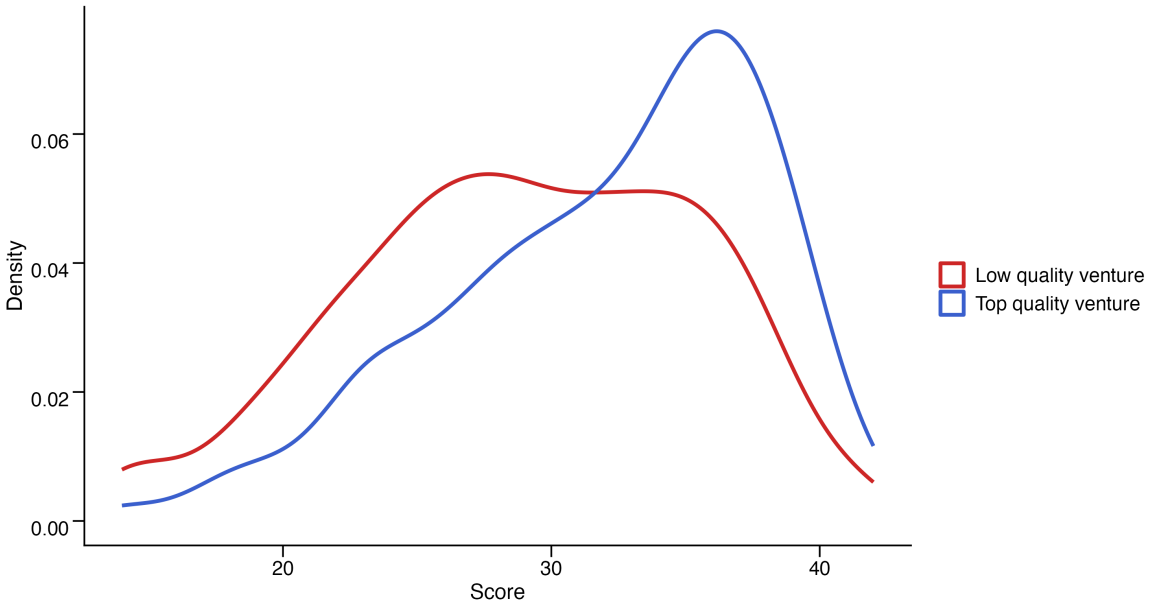
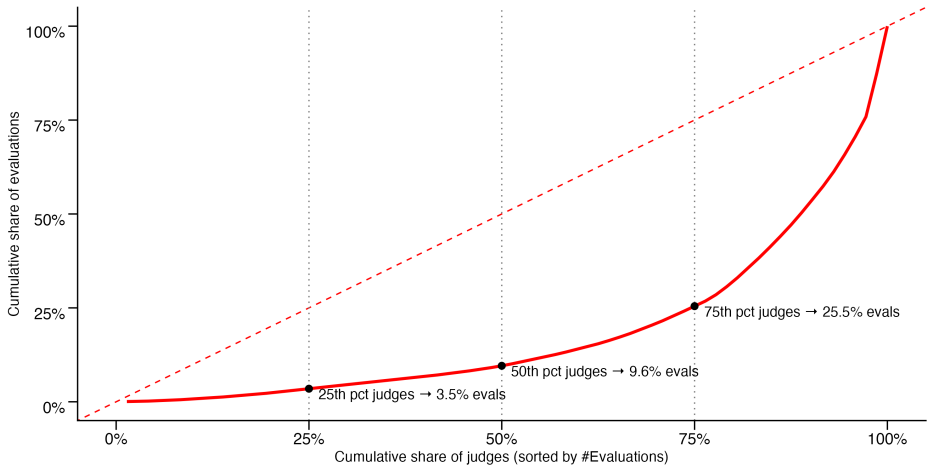
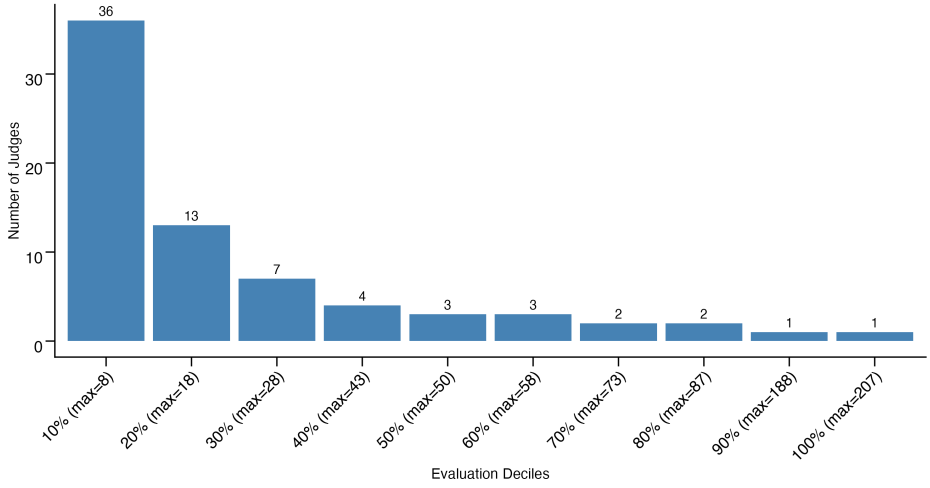


Figure K1: Score and Venture type
The figure shows the distributions of the score conditional on the underlying value of ventures in the entire sample of 1638 evaluations.



(a) Lorenz Curve



(b) Number of Judges by experience

Figure K2: Judges' Experience

The figures illustrate the distribution of evaluation activity across judges. Panel (a) plots the cumulative share of evaluations as a function of the cumulative share of judges, sorted by experience. Panel (b) reports the number of judges in each decile of the distribution of evaluations per judge, where each decile corresponds to successive shares of total evaluations, ordering judges by the number of evaluations performed. Overall, the results show that a large number of judges conduct only a few evaluations, while a small group of highly active judges account for the majority of evaluations.

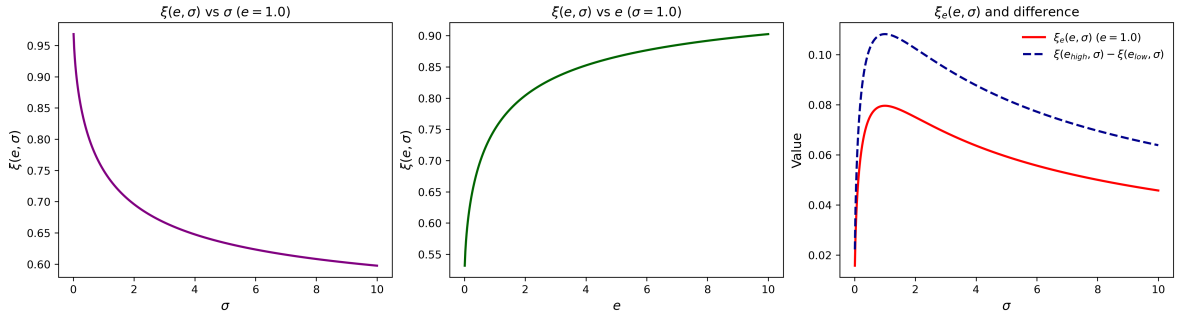


Figure K3: Expected accuracy as a function of expertise and complexity. The numerical simulation is derived from:

$$\xi(\sigma, e) = \frac{1}{2} + \frac{1}{\pi} \arctan \sqrt{\frac{e}{\sigma}}$$

The figure shows how expected accuracy varies with complexity, expertise, and their interaction. In the left panel, expected accuracy is plotted as a function of complexity. In the center panel, expected accuracy is shown as a function of expertise. The right panel reports the marginal value of expertise as a function of complexity: the partial derivative of expected accuracy with respect to expertise.

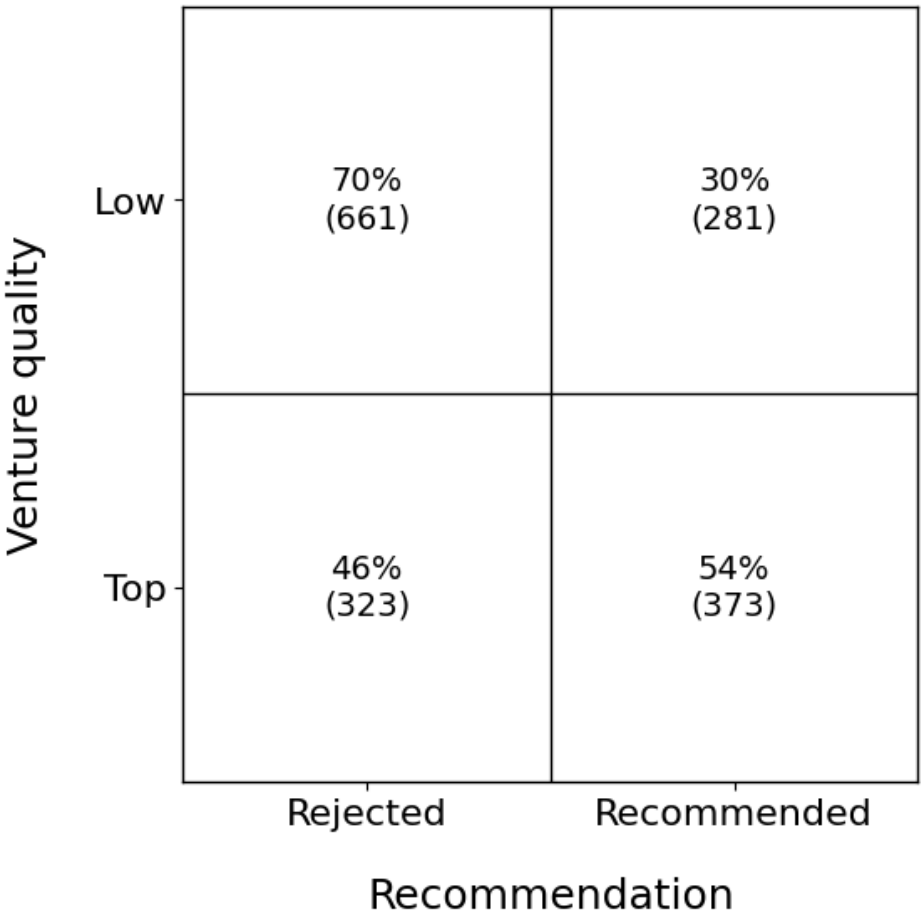


Figure K4: Judges' Prediction Accuracy
The confusion matrix compares judges' admission recommendations with ventures' latent economic quality. Each row sums to 100%, reporting the proportion of evaluations of ventures of a given quality (rows) that resulted in either rejection or recommendation (columns), with absolute counts shown in parentheses. For example, among top-quality ventures, 54% were recommended and 46% were rejected, while among low-quality ventures, 30% were recommended and 70% were rejected.

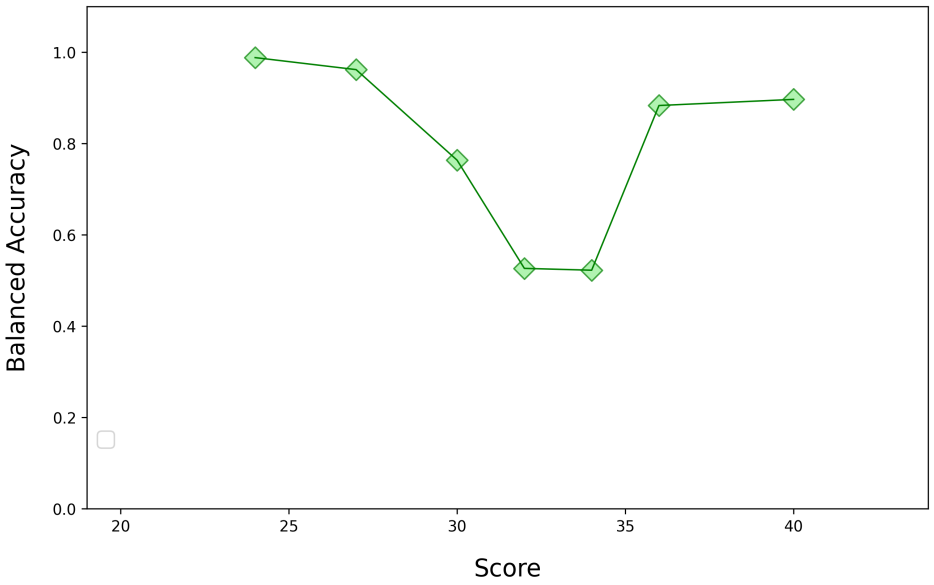


Figure K5: Predicting Judges' recommendations
The figure shows the balanced accuracy rates for the estimated Bayes classifier replicating judges' classification assessments of ventures as top quality in their batches. Accuracy rates were computed by grouping evaluations into eight bins based on their total criteria grades. Each bin contains between 11% and 14% of the 1638 evaluations. For each bin, the balanced accuracy rate was calculated as the weighted average of two metrics: the ratio of evaluations where the model correctly predicts a judge's classification of a venture as a top performer, relative to the total number of evaluations where judges classified ventures as top quality (true positive rate), and the ratio of evaluations where the model correctly predicts a judge's classification of a venture as a low performer, relative to the total number of evaluations where judges classified ventures as low quality (true negative rate).

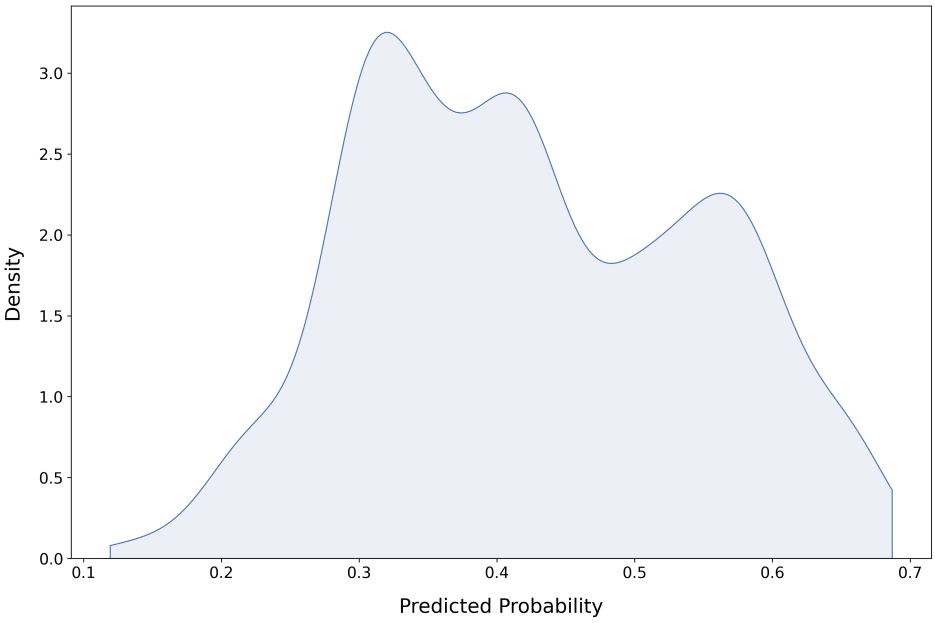


Figure K6: Predicted Probabilities by Bayesian Model
The figure illustrates the distribution of predicted probabilities of observing a top quality venture that estimated Bayes classifier generated from the score of evaluations.

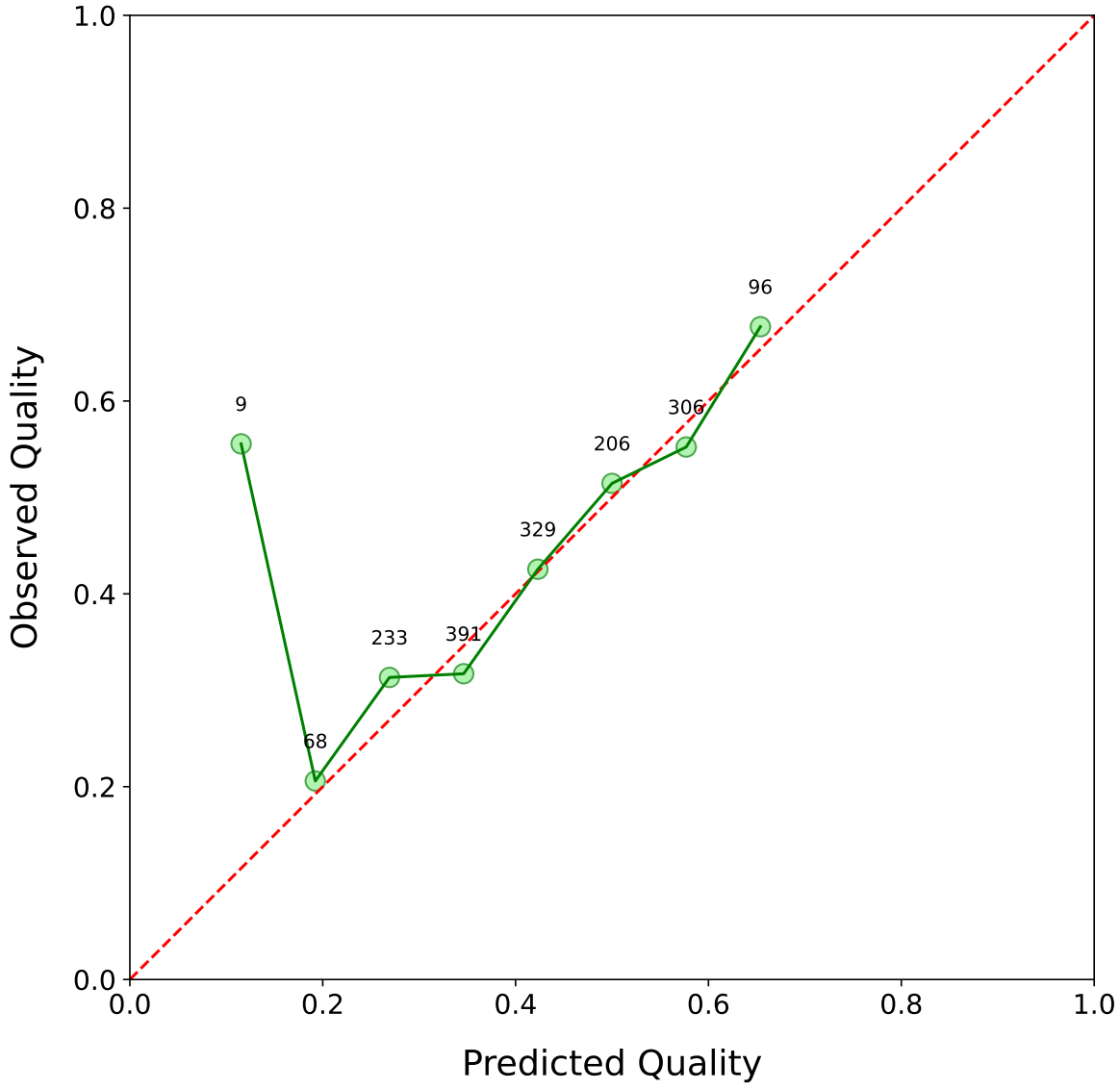


Figure K7: Calibration Plot of the Bayesian Model
The figure shows the number of evaluations in the evaluations set for which the estimated Bayesian classifier assigned a given probability of belonging to a top-quality venture. It compares these probabilities to the actual frequency of observed top-quality ventures among evaluations with the corresponding assigned probability. To compute this, we divided the predicted probabilities into 8 intervals: $[8.3\%, 16.7\%), [16.7\%, 25\%), ..., [66.7\%, 75\%)$, and calculated the frequency of evaluations that belonged to top-quality ventures and had predicted probabilities within each interval. For example, figure shows there were 329 evaluations for which the Bayes classifier assigned a probability between 38.5% and 46.2% of belonging to a top-quality venture, with an observed frequency of top-quality ventures of approximately 42.6%.

L. Additional Tables

			Survival		FTE		€Mln Funding	
Batch	Evaluations	$P(\hat{a}(\hat{s})=1)$	Mean-diff	r_{pb}	Mean-diff	r_{pb}	Mean-diff	r_{pb}
2021 Fall	144	35%	15%*	0.14*	7.22***	0.29***	0.36	0.02
2021 Summer	156	40%	29%***	0.3***	7.55***	0.29***	3.77*	0.19**
2022 Spring	145	40%	14%**	0.19**	3.08***	0.29***	0.05	0.05
2022 Summer	283	36%	33%***	0.32***	5.33***	0.26***	0.33***	0.24***
2022 Winter	156	37%	17%**	0.18**	8.25***	0.31***	0.92**	0.21***
2023 Spring	219	47%	0%	0	1.85**	0.16**	0	0.01
2023 Summer	188	41%	14%**	0.16**	1.69***	0.25***	0.09**	0.17**
2023 Winter	154	40%	8%	0.1	2.24***	0.24***	0.06	0.11
2024 Winter	193	44%	9%	0.1	1**	0.16**	0.02	0.02
Overall	1638	40%	17%***	0.18***	3.87***	0.22***	0.51**	0.06**

Table L1: Outcomes differences by recommendation status
The table reports statistics by batch and for the overall sample, comparing ventures that received a positive recommendation with those that were rejected. Column (2) shows the number of evaluations in each batch. Column (3) reports the proportion of applications that received a positive recommendation. Columns (4), (6), and (8) present the difference in means between recommended and rejected applications for three key business outcomes: the probability of survival, the number of full-time equivalents (FTEs), and the amount of funding raised post-application, respectively. Columns (5), (7), and (9) report the point-biserial correlation coefficients between judges' recommendation decisions and these same business outcomes. Asterisks indicate statistical significance: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Metric	Value	Interpretation of the value
Balanced Accuracy	60%	Moderate discrimination of ventures quality; close to judges' accuracy (62%).
Brier Score (BS)	0.233	Probabilistic error modestly below the constant-rate classification (0.244).
Uncertainty	0.244	High variance in economic quality consistent with base rate $\bar{v} \approx 0.42$.
Resolution	0.014	Predictions cluster near the base rate, limiting separation of economic quality conditional on score.
Reliability	0.001	Very small miscalibration; predicted probabilities track observed frequencies closely.
Brier Skill Score (BSS)	4.8%	Limited improvement over classification based on base rate.
Expected Calibration Error (ECE)	2.1%	Very small average binwise calibration gap.

Table L2: Estimated Bayes Classifier
Performance diagnostics for the estimated Bayes classifier on the full sample of evaluations. The table reports the Balanced Accuracy, Brier score (BS), the Brier Skill Score (BSS), and the Expected Calibration Error (ECE). The Brier score is the mean squared prediction error, $\mathrm{BS} = \mathbb{E}[(\hat{v} - \hat{\eta}(\hat{s}))^2]$, and admits the decomposition

$$\mathrm{BS} = \underbrace{\bar{\eta}(1-\bar{\eta})}_{\text{Uncertainty}} - \underbrace{\mathbb{E}\left[(m(P)-\bar{\eta})^2\right]}_{\text{Resolution}} + \underbrace{\mathbb{E}\left[(m(P)-P)^2\right]}_{\text{Reliability}},$$

where $\bar{\eta} = \mathbb{E}[\hat{v}]$ is the base rate, $P = \hat{\eta}(\hat{s})$ is the predicted posterior probability, and $m(P) = \mathbb{E}[\hat{v} \mid \hat{\eta}]$ is the calibration curve. The Brier Skill Score is defined as $\mathrm{BSS} = 1 - \mathrm{BS}/[\bar{\eta}(1-\bar{\eta})]$ and measures predictive performance relative to a baseline forecast that always predicts the base rate. The Expected Calibration Error is defined as $\mathrm{ECE} = \mathbb{E}[|m(P) - P|]$ and measures the average absolute deviation between predicted probabilities and realized frequencies.

Dependent Variable:	Accuracy					
	Disagreement			Text-based		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	63.70***	91.11***		69.90***	92.87***	
	(4.598)	(11.89)		(3.791)	(10.98)	
Complexity$_i$, Q2	3.291	3.469	4.889	-13.59***	-12.83***	-12.29**
	(5.488)	(5.497)	(5.696)	(4.851)	(4.817)	(4.842)
Complexity$_i$, Q3	-6.679	-6.617	-3.351	-11.69**	-11.07**	-9.696**
	(7.043)	(7.101)	(7.183)	(4.879)	(4.836)	(4.504)
Complexity$_i$, Q4	-4.520	-5.940	-3.099	-9.403*	-8.006	-8.344
	(5.293)	(5.340)	(5.468)	(5.546)	(5.356)	(5.157)
High Experience$_j$	8.899**	8.544*	5.445	-1.122	-1.892	-6.755*
	(4.397)	(4.398)	(4.199)	(4.276)	(4.448)	(3.852)
High Experience$_j$ × Complexity$_i$, Q2	-10.80*	-11.32*	-12.01**	11.79**	12.40**	12.87**
	(5.901)	(5.684)	(5.751)	(5.842)	(5.778)	(6.290)
High Experience$_j$ × Complexity$_i$, Q3	-5.411	-6.228	-7.227	1.669	2.460	3.320
	(7.862)	(7.800)	(7.589)	(4.576)	(4.404)	(4.301)
High Experience$_j$ × Complexity$_i$, Q4	-7.120	-7.866	-8.756	5.949	4.920	5.254
	(6.321)	(6.083)	(6.316)	(5.971)	(5.665)	(5.350)
Fixed-effects						
Batch (9)			Yes			Yes
Controls: $\hat{s}_{i,j}$		Yes	Yes		Yes	Yes
Fit statistics						
Observations	1,638	1,638	1,638	1,638	1,638	1,638

Table L3: Accuracy Gain for Experienced Judges by Venture Complexity
This table reports estimates from the regression:

$$\begin{aligned}\text{Accuracy}_{i,j} = \alpha + \beta_1 \text{High Experience}_j + \sum_{f=2}^{4} \beta_f \hat{\sigma}_i^f \\ + \sum_{f=2}^{4} \gamma_f \text{High Experience}_j \hat{\sigma}_i^f \\ + \hat{s}_{i,j} + \delta_{t(i)} + \varepsilon_{i,j}\end{aligned}$$

where $\hat{\sigma}_i^f$ are indicator variables for the 2nd, 3rd, and 4th quartiles of venture complexity, *High Experience* equals 1 for judges above the median of the distribution of the number of cumulative evaluations, $\hat{s}_{i,j}$ are evaluations' scores, $\delta_{t(i)}$ batch fixed effects . Columns (1)–(3) report estimates from OLS regressions using the disagreement based complexity measure; columns (4)-(6) use the text based complexity measure RIX. Standard errors are clustered at the venture and evaluator level. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	$\gamma_L = 0\%$	$\gamma_L = 10\%$	$\gamma_L = 20\%$	$\gamma_L = 30\%$
$\gamma_S = 0\%$				
$\gamma_K = 0\%$	0% (0%)	3% (2%)	3% (2%)	5% (6%)
$\gamma_K = 10\%$	6% (12%)	7% (15%)	7% (15%)	6% (18%)
$\gamma_K = 20\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)
$\gamma_K = 30\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)
$\gamma_S = 10\%$				
$\gamma_K = 0\%$	3% (3%)	5% (4%)	5% (4%)	6% (6%)
$\gamma_K = 10\%$	6% (12%)	7% (15%)	7% (15%)	6% (18%)
$\gamma_K = 20\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)
$\gamma_K = 30\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)
$\gamma_S = 20\%$				
$\gamma_K = 0\%$	0% (0%)	3% (2%)	3% (2%)	4% (6%)
$\gamma_K = 10\%$	6% (12%)	7% (15%)	7% (15%)	6% (17%)
$\gamma_K = 20\%$	5% (14%)	6% (16%)	6% (16%)	6% (18%)
$\gamma_K = 30\%$	5% (14%)	6% (16%)	6% (16%)	6% (18%)
$\gamma_S = 30\%$				
$\gamma_K = 0\%$	3% (3%)	5% (4%)	5% (4%)	6% (6%)
$\gamma_K = 10\%$	6% (12%)	7% (15%)	7% (15%)	6% (18%)
$\gamma_K = 20\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)
$\gamma_K = 30\%$	6% (14%)	7% (16%)	7% (16%)	7% (18%)

Table L4: Simulated treatment effect and firms class
The table reports the proportion of ventures that change classification as top-quality or not, based on the simulated standardized treatment effect sizes for survival γ_S, employment γ_L and raised capital γ_K. Values in parenthesis denote the fraction of changes occurring among admitted ventures only.

Variable	Min	Mean	Median	Max	No effect
Judges accuracy	56%	57%	57%	63%	61%
Estimated Bayes, Accuracy $\hat{v}$	56%	57%	57%	62%	62%
Estimated Bayes, Accuracy $\hat{a}(\hat{s})$	75%	77%	75%	80%	79%

Table L5: Accuracy rates across simulations
This table presents summary statistics for the accuracy rates of different models across various simulation scenarios, alongside the accuracy rate for the baseline scenario with no treatment effect. The models compared include judges' accuracy in predicting $\hat{v}$, as well as the estimated Bayes predicting both $\hat{a}(\hat{s})$ and $\hat{v}$